

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2002-163248

(43)Date of publication of application : 07.06.2002

(51)Int.Cl.

G06F 17/21

G06F 12/00

G06F 17/30

(21)Application number : 2000-357568

(71)Applicant : FUJITSU LTD

(22)Date of filing : 24.11.2000

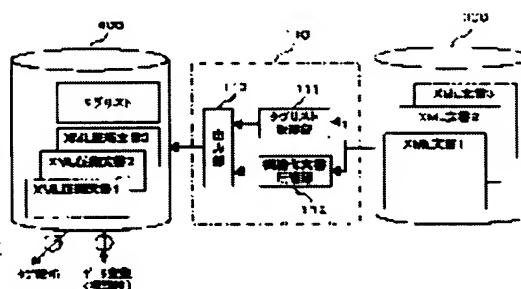
(72)Inventor : SATO NOBUKO

(54) STRUCTURED DOCUMENT COMPRESSOR, STRUCTURED DOCUMENT RESTORING DEVICE AND STRUCTURED DOCUMENT PROCESSING SYSTEM

(57)Abstract:

PROBLEM TO BE SOLVED: To reduce the data amount of a structured document by compressing the structured document while keeping the visibility, flexibility and expandability of the data structure which are advantages of the structure document, and to reduce load of tag analysis by dispensing with wasteful tag analysis in handling a number of structured documents having the same data structure.

SOLUTION: This structured document processing system is provided with a tag list generating part 111 for generating one tag list by extracting tags in the structured documents in the appearing order to make a list, a structured document compressing part 112 for generating a compressed document by replacing a tag in each structure document with a designated partition code, and an output part 113 for making the above one tag list generated by the tag list generating part 111 correspond to a plurality of compressed documents generated by the structured document compressing part 112 concerning each of the plurality of structured documents and outputting the same as the compression result.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

Copyright (C); 1998,2003 Japan Patent Office

(19)日本国特許庁 (J P)

(12) 公 開 特 許 公 報 (A)

(11)特許出願公開番号
特開2002-163248
(P2002-163248A)

(43)公開日 平成14年6月7日(2002.6.7)

(51)Int.Cl. ⁷	識別記号	F I	テーマコード*(参考)
G 0 6 F 17/21	5 0 1 5 7 0	G 0 6 F 17/21	5 0 1 T 5 B 0 0 9 5 7 0 G 5 B 0 7 5 5 7 0 D 5 B 0 8 2
12/00	5 1 1	12/00	5 1 1 A
17/30	2 3 0	17/30	2 3 0 A
審査請求 未請求 請求項の数 5 O L (全 25 頁)			

(21)出願番号 特願2000-357568(P2000-357568)

(22)出願日 平成12年11月24日(2000.11.24)

(71)出願人 000005223

富士通株式会社

神奈川県川崎市中原区上小田中4丁目1番
1号

(72)発明者 佐藤 宣子

神奈川県川崎市中原区上小田中4丁目1番
1号 富士通株式会社内

(74)代理人 100092978

弁理士 真田 有

Fターム(参考) 5B009 NA05 SA08

5B075 NR03 NR16

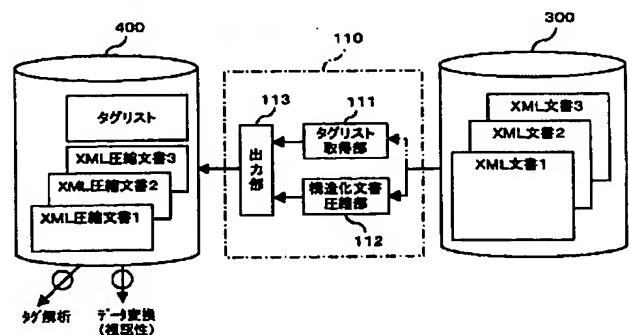
5B082 GA01

(54)【発明の名称】 構造化文書圧縮装置および構造化文書復元装置並びに構造化文書処理システム

(57)【要約】

【課題】構造化文書の利点であるデータ構造の視認性や柔軟性／拡張性の高さを生かしたまま構造化文書を圧縮できるようにして、構造化文書のデータ量の削減をはかるとともに、同一のデータ構造を有する多数の構造化文書を取り扱う際に無駄なタグ解析を行なう必要を一切なくして、タグ解析の負荷の低減をはかる。

【解決手段】構造化文書におけるタグを出現順序に従って抽出してリスト化した一つのタグリストを生成するタグリスト生成部111と、各構造化文書中のタグを所定の区切りコードに置き換えた圧縮文書を生成する構造化文書圧縮部112と、タグリスト生成部111により生成された前記一つのタグリストと複数の構造化文書のそれぞれについて構造化文書圧縮部112により生成された複数の圧縮文書とを対応させ圧縮結果として出力する出力部113とをそなえる。



【特許請求の範囲】

【請求項 1】 同一のデータ構造を有する複数の構造化文書を圧縮する装置であって、
該構造化文書におけるタグを出現順序に従って抽出してリスト化した、該複数の構造化文書について共通の一つのタグリストを取得するタグリスト取得部と、
各構造化文書中のタグを所定の区切りコードに置き換えた圧縮文書を生成する構造化文書圧縮部と、
該タグリスト取得部により取得された前記一つのタグリストと、該複数の構造化文書のそれぞれについて該構造化文書圧縮部により生成された複数の圧縮文書とを対応させ該複数の構造化文書の圧縮結果として出力する出力部とをそなえたことを特徴とする、構造化文書圧縮装置。

【請求項 2】 構造化文書を圧縮する装置であって、
該構造化文書中のタグを検出するタグ検出部と、
該タグ検出部により検出された該タグを所定の区切りコードに置き換えて圧縮するタグ圧縮部とをそなえたことを特徴とする、構造化文書圧縮装置。

【請求項 3】 同一のデータ構造を有する複数の構造化文書中のタグを所定の区切りコードに置き換えることにより生成された複数の圧縮文書を、該複数の構造化文書におけるタグを出現順序に従ってリスト化したタグリストに基づいて復元する装置であって、
該タグリストに対応するデータ構造をメモリ上に複製データ構造として展開・複製する複製部と、
該複製データ構造におけるタグの位置と各圧縮文書中の前記所定の区切りコードの位置とを対応させながら、各圧縮文書中の要素内容を該メモリ上における該複製データ構造の所定領域に書き出す書出部とをそなえたことを特徴とする、構造化文書復元装置。

【請求項 4】 構造化文書中のタグを所定の区切りコードに置き換えることにより生成された圧縮文書を復元する装置であって、
該構造化文書におけるタグを出現順序に従ってリスト化したタグリストを予め保持するタグリスト保持部と、
該圧縮文書中の前記所定の区切りコードを検出する区切りコード検出部と、
該タグリストにおけるタグの位置と該区切りコード検出部により検出された前記所定の区切りコードの位置とを対応させながら、該区切りコード検出部により検出された前記所定の区切りコードを、該タグリストにおける対応するタグに置き換えて復元するタグ復元部とをそなえたことを特徴とする、構造化文書復元装置。

【請求項 5】 同一のデータ構造を有する複数の構造化文書に対する処理を行なうべく、該複数の構造化文書を圧縮する構造化文書圧縮装置と、該構造化文書圧縮装置による圧縮データを該複数の構造化文書に復元する構造化文書復元装置とを含んで構成される構造化文書処理システムにおいて、

該構造化文書圧縮装置が、
該構造化文書におけるタグを出現順序に従って抽出してリスト化した、該複数の構造化文書について共通の一つのタグリストを取得するタグリスト取得部と、
各構造化文書中のタグを所定の区切りコードに置き換えた圧縮文書を生成する構造化文書圧縮部と、
該タグリスト取得部により取得された前記一つのタグリストと、該複数の構造化文書のそれぞれについて該構造化文書圧縮部により生成された複数の圧縮文書とを対応させ該複数の構造化文書の圧縮結果として出力する出力部とをそなえて構成されるときに、
該構造化文書復元装置が、
該複数の圧縮文書の復元結果を格納するメモリと、
該タグリストに対応するデータ構造を該メモリ上に複製データ構造として展開・複製する複製部と、
該複製データ構造におけるタグの位置と各圧縮文書中の前記所定の区切りコードの位置とを対応させながら、各圧縮文書中の要素内容を該メモリ上における該複製データ構造の所定領域に書き出す書出部とをそなえて構成されたことを特徴とする、構造化文書処理システム。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】 本発明は、SGML (Standard Generalized Markup Language) や XML (eXtensible Markup Language) 等の構造化文書のための圧縮・復元に係る技術に関し、特に、同一のデータ構造を有する多数の構造化文書、例えば同一フォーマットの多数の伝票類を取り扱う際に用いて好適の、構造化文書圧縮装置および構造化文書復元装置並びに構造化文書処理システムに関する。

【0002】

【従来の技術】 近年、文字コード、画像データ等の様々な種類のデータがコンピュータで扱われるようになるに伴い、取り扱われるデータ量も増大している。そのような大量のデータをそのまま取り扱うと、そのデータを記憶するために多大な記憶容量が必要になり、そのデータの伝送に多大な時間を要することになるが、そのデータ中の冗長な部分を省いて圧縮すれば、記憶容量を減らしたり、遠隔地への伝送を高速化したりすることができる。様々なデータを一つの方式で圧縮することができる方法として、例えばユニバーサル符号化が提案されている。

【0003】 一般的なユニバーサルデータ圧縮方式としては、データ系列の類似性を利用した辞書型符号化方式と、データ列の出現頻度を利用した確率統計型符号化方式とがある（例えば、CQ 出版社刊：植松友彦著 “文書データ圧縮アルゴリズム入門” 参照）。いずれの方式においても、生成される圧縮データは、バイナリコード

（例えば、8 ビットコードで見ると、0x00～0xFF の全てのコードを使用する）になっている。例えば図

3

19は一般的なユニバーサルデータ圧縮について説明するための図であるが、この図19に示すように、ユニバーサルデータ圧縮により、XML文書1, 2, 3はそれぞれバイナリデータ1, 2, 3に圧縮・変換されることになる。

【0004】一方、最近では計算機で取り扱うデータの形式を統一する動きがある。即ち、これまで、計算機やアプリケーションによってバラバラであったデータ形式を、異なる計算機やアプリケーションでも使用できるようにするものである。このようなデータ形式を統一するための規格として、XMLが、1998年2月にW3C (World Wide Web Consortium)によって正式に勧告されている。このXMLは、同様の規格SGMLのサブセットになっており、文書自身の中にタグを埋め込む形で、その文書のデータ構造が記述される。XMLやSGMLにより記述された文書は、一般に構造化文書と呼ばれる。

【0005】このような構造化文書は、データ構造をタグとして文書に埋め込んだ構成を採用しているため、データ構造について高い柔軟性／拡張性を有するという利点を得られる。また、人が見て意味のあるテキストによってタグが記述されているので、XML文書のデータ構造の視認性が高い。従って、データ交換を行ないやすくなり、複数種類の記述方法を緩やかに統合することができるといふ利点も得られる。このことは、構造化文書を成すデータが、アプリケーションに依存しないことを意味する。

【0006】以下では、XML規格に従って、「<」と「>」とで囲まれた文字列（要素名を表す文字列）をタグ、「<文字列>」を開始タグ、「</文字列>」を終了タグ、これらの開始タグと終了タグとの間にはさまれた領域に記述された文字列を要素（もしくは要素内容）と呼ぶ。

【0007】XML規格に従って記述されたXML文書の利用は、ウェブ(Web)やデータベースの分野を中心に増えつつあり、特に、XML文書は、EDI (Electric Data Interchange), EC (Electric Commerce), 携帯電話サービス, デジタルテレビ向けサービス, Webサービスなどで広く利用されつつある。図20は、上述のようなXML文書を取り扱う一般的なシステムの構成例を示すブロック図である。この図20に示すシステムでは、ハードディスク（データベース）10、読出部20、メモリ展開部30およびメモリ40がそなえられている。

【0008】ハードディスク（データベース）10は、XML文書（図20では3つのXML文書1, 2, 3）を格納・保持するものであり、読出部20は、ハードディスク10のXML文書をメモリ展開部30へ読み出すものである。メモリ展開部30は、読出部20から入力されたXML文書を、メモリ40上に展開するためのも

4

ので、解析部31、生成部32および格納部33を有して構成されている。

【0009】ここで、解析部31は、メモリ40上に展開すべき各XML文書におけるタグを解析することにより、各XML文書のデータ構造（木構造）を解析するものであり、生成部32は、解析部31によるデータ構造解析結果に従って、各XML文書に応じた文書データを生成するものであり、格納部33は、生成部32により生成された文書データをメモリ40上に展開し格納するものである。

【0010】上述の構成により、図20に示すシステムでは、ハードディスク10に格納されているXML文書が、読出部20により読み出されてメモリ展開部30に入力されると、解析部31により、そのXML文書のデータ構造（木構造）が解析される。そして、生成部32において、解析部31によるデータ構造解析結果に従い、そのXML文書に応じた文書データが生成され、生成された文書データが、格納部33によりメモリ40上に展開されて格納される。

【0011】

【発明が解決しようとする課題】ところで、XML文書（構造化文書）は、データ構造について高い柔軟性／拡張性を有するとともに高い視認性を有するが、人が見て意味を理解できるようにタグを記述するため、冗長な記述となり、そのXML文書のデータ量は大幅に増大する。従って、XML文書を圧縮して、データ量を削減することが望まれている。

【0012】しかしながら、ユニバーサルデータ圧縮を用いると、例えば図19に示すごとく、圧縮データが全てバイナリデータになるため、XML文書の利点の一つであるデータの視認性が全くなくなってしまう、圧縮データを元の状態に復元しなければ、その圧縮データの内容を一切把握することができない。このため、圧縮データの復元アルゴリズムが分からなければ、データ交換もタグ解析も行なうことができない。

【0013】一方、図20を参照しながら説明したごとくXML文書をメモリ40上に展開する際、XML文書（構造化文書）ではデータ構造がタグとしてXML文書中に埋め込まれているため、解析部31によるタグ解析処理（データ構造解析処理）の負荷が高くなる。特に、同一のデータ構造を有する多数のXML文書（例えば発注伝票）をメモリ40上に展開する場合、全てのXML文書が同一のデータ構造を有しているにもかかわらず、XML文書一つ一つについてデータ構造解析処理を行なわなければならない、無駄なタグ解析処理の量が増大し、その処理による負荷が極めて高くなるという課題があった。

【0014】本発明は、このような課題に鑑み創案されたもので、構造化文書の利点であるデータ構造の視認性や柔軟性／拡張性の高さを生かしたまま、構造化文書を

圧縮できるようにして、構造化文書のデータ量の削減をはかるとともに、同一のデータ構造を有する多数の構造化文書を取り扱う際に無駄なタグ解析を行なう必要を一切なくして、タグ解析の負荷の低減をはかった、構造化文書圧縮装置および構造化文書復元装置並びに構造化文書処理システムを提供することを目的とする。

【0015】

【課題を解決するための手段】上記目的を達成するために、本発明の構造化文書圧縮装置（請求項1）は、同一のデータ構造を有する複数の構造化文書を圧縮する装置であって、該構造化文書におけるタグを出現順序に従って抽出してリスト化した該複数の構造化文書について共通の一つのタグリストを取得するタグリスト取得部と、各構造化文書中のタグを所定の区切りコードに置き換えた圧縮文書を生成する構造化文書圧縮部と、該タグリスト生成部により生成された前記一つのタグリストと、該複数の構造化文書のそれぞれについて該構造化文書圧縮部により生成された複数の圧縮文書とを対応させ該複数の構造化文書の圧縮結果として出力する出力部とをそなえたことを特徴としている。

【0016】また、本発明の構造化文書復元装置（請求項3）は、上述した本発明の構造化文書圧縮装置（請求項1）により生成された複数の圧縮文書を復元する装置であって、タグリストに対応するデータ構造をメモリ上に複製データ構造として展開・複製する複製部と、その複製データ構造におけるタグの位置と各圧縮文書中の前記所定の区切りコードの位置とを対応させながら各圧縮文書中の要素内容を該メモリ上における該複製データ構造の所定領域に書き出す書出部とをそなえたことを特徴としている。

【0017】そして、本発明の構造化文書処理システム（請求項5）は、同一のデータ構造を有する複数の構造化文書に対する処理を行なうべく、上述した本発明の構造化文書圧縮装置（請求項1）および構造化文書復元装置（請求項3）を含んで構成されたことを特徴としている。

【0018】上述した、本発明の構造化文書圧縮装置（請求項1）、あるいは、本発明の構造化文書処理システム（請求項5）における構造化文書圧縮装置では、複数の構造化文書について共通のデータ構造が、タグリスト取得部により、一つのタグリストとして取得されるとともに、各構造化文書は、構造化文書圧縮部のタグ圧縮処理（タグを所定の区切りコードに置き換える処理）により圧縮文書に変換された後、一つのタグリストと複数の圧縮文書とが、複数の構造化文書についての圧縮結果として出力部から出力される。

【0019】従って、複数の構造化文書の圧縮結果は、データ構造（一つのタグリスト）とデータ内容（複数の圧縮文書）とに分離された状態で出力される。また、上述のごとく生成された圧縮文書では、タグが所定の区切

りコードに置換されているだけで、データ内容（要素内容）はそのまま記述されている。このため、構造化文書の利点であるデータ構造の視認性や柔軟性／拡張性の高さを生かしたまま、構造化文書を圧縮することができる。

【0020】さらに、上述した、本発明の構造化文書復元装置（請求項3）、あるいは、本発明の構造化文書処理システム（請求項5）における構造化文書復元装置では、タグリストに対応するデータ構造が、複製部により、圧縮文書毎にメモリ上に複製データ構造として展開・複製される。そして、各圧縮文書中の要素内容が、書出部により、複製データ構造におけるタグの位置と各圧縮文書中の所定の区切りコードの位置との対応をとりながら、メモリ上における複製データ構造の所定領域に書き出される。これにより、各圧縮文書（各構造化文書）が、メモリ上に展開された状態で復元されることになる。

【0021】このとき、複数の構造化文書の圧縮結果は、データ構造（一つのタグリスト）とデータ内容（複数の圧縮文書）とに分離されているので、一つのタグリストに対する解析処理を一度だけ行なって、複数の圧縮文書に共通のデータ構造を取得してしまえば、後は、取得されたデータ構造を複製して流用することにより、圧縮文書毎に一々タグ解析を行なう必要をなくすることができる。

【0022】一方、本発明の構造化文書圧縮装置（請求項2）は、例えば上述した構造化文書圧縮装置（請求項1）における構造化文書圧縮部として用いられるものであって、構造化文書を圧縮すべく、該構造化文書中のタグを検出するタグ検出部と、該タグ検出部により検出された該タグを所定の区切りコードに置き換えて圧縮するタグ圧縮部とをそなえたことを特徴としている。

【0023】また、本発明の構造化文書復元装置（請求項4）は、上述した本発明の構造化文書圧縮装置（請求項2）により生成された圧縮文書を復元する装置であって、該構造化文書におけるタグを出現順序に従ってリスト化したタグリストを予め保持するタグリスト保持部と、該圧縮文書中の所定の区切りコードを検出する区切りコード検出部と、該タグリストにおけるタグの位置と該区切りコード検出部により検出された前記所定の区切りコードの位置とを対応させながら、該区切りコード検出部により検出された前記所定の区切りコードを、該タグリストにおける対応するタグに置き換えて復元するタグ復元部とをそなえたことを特徴としている。

【0024】上述した本発明の構造化文書圧縮装置（請求項2）では、構造化文書中のタグがタグ検出部により検出されると、そのタグは、タグ圧縮部により所定の区切りコードに置き換えられて圧縮される。このような単純な置換処理によって圧縮文書が生成される。そして、生成された圧縮文書では、上述した通り、タグが所定の

区切りコードに置換されているだけで、データ内容（要素内容）はそのまま記述されている。従って、構造化文書の利点であるデータ構造の視認性や柔軟性／拡張性の高さを生かしたまま、構造化文書を圧縮することができる。

【0025】また、上述した本発明の構造化文書復元装置（請求項4）では、復元対象の圧縮文書に対応したタグリストが、タグリスト保持部に予め保持されており、圧縮文書中の区切りコードが区切りコード検出部によって検出されると、その区切りコードは、タグ復元部により、その区切りコードに対応したタグに置き換えられる。

【0026】このとき、タグ復元部においては、タグリストにおけるタグの位置と区切りコード検出部により検出された所定の区切りコードの位置との対応をとることにより、検出された所定の区切りコードがタグリスト中のどのタグに対応するかが認識されるので、所定の区切りコードを適切なタグに復元することができる。これにより、圧縮文書は、元の構造化文書に復元される。

【0027】

【発明の実施の形態】以下、図面を参照して本発明の実施の形態を説明する。

〔1〕第1実施形態の説明

図1は本発明の第1実施形態における構造化文書圧縮装置の構成を示すブロック図、図2は本発明の第1実施形態における構造化文書復元装置（メモリ展開部）の構成を示すブロック図である。

【0028】図1に示す構造化文書圧縮装置110および図2に示すメモリ展開部（構造化文書復元装置）210は、同一のデータ構造（文書構造、木構造）を有する複数の構造化文書に対する処理を行なうための構造化文書処理システムに含まれ、この構造化文書処理システムの一部を構成するものである。

【0029】なお、以下に説明する各実施形態においては、構造化文書がXML（eXtensible Markup Language）であり、特に、同一のデータ構造を有する多数のXML文書を取り扱う場合について説明する。また、構造化文書処理システムによって実行される処理は、例えばXML文書の蓄積、加工、転送等である。なお、処理対象となるXML文書は、例えば図3（A）を参照しながら後述するような発注伝票文書である。

【0030】そして、構造化文書処理システムにおいては、XML文書のための記憶容量を削減したり前記処理を高速化すべくXML文書を圧縮するための構造化文書圧縮装置110がそなえられ、さらに、構造化文書圧縮装置110による圧縮データをメモリ214上に復元・展開するためのメモリ展開部（構造化文書復元装置）210がそなえられている。

【0031】第1実施形態の構造化文書圧縮装置110は、図1に示すように、タグリスト取得部111、構造

化文書圧縮部112および出力部113をそなえて構成される一方、第1実施形態のメモリ展開部210は、図2に示すように、解析部211、複製部212および書出部213をそなえて構成されている。

【0032】ここで、構造化文書圧縮装置110およびメモリ展開部210は、同一のコンピュータ上、もしくは、それぞれ異なるコンピュータ上にそなえられている。そして、構造化文書圧縮装置110を成すタグリスト取得部111、構造化文書圧縮部112および出力部113としての機能は、コンピュータ上で所定のプログラム（構造化文書圧縮プログラム）を実行することにより実現される。同様に、メモリ展開部210を成す解析部211、複製部212および書出部213としての機能は、コンピュータ上で所定のプログラム（構造化文書復元プログラム）を実行することにより実現されるようになっている。

【0033】さて、図1において、ハードディスク（データベース）300は、圧縮対象である、同一データ構造を有する複数のXML文書（図1ではXML文書1～3）を予め格納・保持するものである。また、ハードディスク（データベース）400は、構造化文書圧縮装置110による圧縮結果を格納・保持するものである。なお、圧縮対象のXML文書と構造化文書圧縮装置110による圧縮結果とを、同じハードディスク300もしくは400に格納するように構成してもよい。

【0034】タグリスト取得部111は、ハードディスク300に格納された複数のXML文書について共通の一つのタグリストを取得するものである。タグリストは、図3（B）を参照しながら後述するごとく、XML文書におけるタグを出現順序に従って抽出してリスト化したもので、どのようなタグがどのような順序で出現するかを示すものである。同一データ構造を有する複数のXML文書についてのタグリストは全て同一のものとなる。

【0035】このタグリスト取得部111は、予め生成されデータベース（図示略）等に格納されているタグリストを、そのデータベースから取得してもよいし、ハードディスク300に格納されている複数のXML文書のうちの少なくとも一つから、タグリストを抽出・生成して取得してもよい。構造化文書圧縮部112は、各XML文書中のタグを所定の区切りコードに置き換えたXML圧縮文書を生成するものである。なお、第1実施形態では、図3（C）を参照しながら後述するごとく、所定の区切りコードとして「，」を用いる。

【0036】出力部113は、タグリスト取得部111により取得された一つのタグリストと、複数のXML文書のそれぞれについて構造化文書圧縮部112により生成された複数のXML圧縮文書（図1ではXML圧縮文書1～3）とを対応させ複数のXML文書の圧縮結果としてハードディスク400に出力・格納するものであ

10

20

30

40

50

る。

【0037】一方、図2において、読出部500は、ハードディスク400に格納された、共通のタグリストと複数のXML圧縮文書とをメモリ展開部210へ読み出すものであり、メモリ展開部210は、タグリストに基づいて複数のXML圧縮文書をメモリ214上に復元・展開すべく、解析部211、複製部212および書出部213を有している。

【0038】解析部211は、ハードディスク400から読出部500によって読み出されたタグリストを解析し、復元・展開対象である複数のXML圧縮文書について共通のデータ構造を解析結果として得るものである。複製部212は、解析部211によって得られた、タグリストに対応するデータ構造を、メモリ214上に複製データ構造として展開・複製するものである。書出部213は、複製データ構造におけるタグの位置と各XML圧縮文書中の区切りコード「,」の位置とを対応させながら、各XML圧縮文書中の要素内容をメモリ214上における複製データ構造の所定領域に書き出すものである。

【0039】次に、上述のごとく構成された、第1実施形態の構造化文書圧縮装置110およびメモリ展開部210の動作について説明する。図1に示す構造化文書圧縮装置110においては、複数のXML文書について共通のデータ構造が、タグリスト取得部111により、一つのタグリストとして取得されるとともに、各XML文書は、構造化文書圧縮部112のタグ圧縮処理により、タグを区切りコード「,」に置き換えたXML圧縮文書に変換される。

【0040】この後、タグリスト取得111により取得されたタグリストと、構造化文書圧縮部112により得られた複数のXML圧縮文書とが、複数のXML文書についての圧縮結果として出力部113から出力され、ハードディスク400に格納される。つまり、第1実施形態では、複数のXML文書の圧縮結果が、データ構造（タグ情報）とデータ内容（タグ情報以外の情報）とに分離された状態で出力されることになる。なお、データ構造（タグ情報）は、前記一つのタグリストであり、データ内容は、区切りコードと要素内容とからなる、複数のXML圧縮文書である。

【0041】このとき、タグリストと各XML圧縮文書とは、例えば図15～図17を参照しながら後述する手法等によって対応付けられており、複数のXML文書について共通のデータ構造を示す一つのタグリストは、複数のXML圧縮文書によって共有される。

【0042】ここで、図3（A）～図3（C）を参照しながら、第1実施形態における具体的なXML文書の圧縮状態について説明する。なお、図3（A）～図3（C）はいずれも第1実施形態におけるデータ例を示すもので、図3（A）はXML文書の一例を示す図、図3

（B）は図3（A）に示すXML文書から得られたタグリストを示す図、図3（C）は図3（A）に示すXML文書の圧縮状態を示す図である。

【0043】図3（A）には、圧縮前つまり圧縮対象のXML文書の一例として、発注伝票をXMLにより記述した例が示されている。この図3（A）に示すXML文書では、開始タグ<発注伝票>、<発注者>、<名前>、<電話番号>、<商品>、<メーカ>、<製品番号>、<製品名>、<価格>と、終了タグ</発注伝票>、</発注者>、</名前>、</電話番号>、</商品>、</メーカ>、</製品番号>、</製品名>、</価格>とにより、XML文書のデータ構造（つまり発注伝票のフォーマット）が定義されている。

【0044】この図3（A）に示すXML文書においては、開始タグ<名前>と終了タグ</名前>との間には、発注者の名前「STUV」が要素内容として記述され、開始タグ<電話番号>と終了タグ</電話番号>との間には、発注者の電話番号「1111」が要素内容として記述され、開始タグ<メーカ>と終了タグ</メーカ>との間には、商品のメーカ「A社」が要素内容として記述され、開始タグ<製品番号>と終了タグ</製品番号>との間には、商品の製品番号「1234」が要素内容として記述され、開始タグ<製品名>と終了タグ</製品名>との間には、商品の製品名「ABCD」が要素内容として記述され、開始タグ<価格>と終了タグ</価格>との間には、商品の価格「980」が要素内容として記述されている。

【0045】また、図3（B）は、図3（A）に示したXML文書のタグリストを示している。このタグリストは、前述した通り、予め何らかの手段により作成されているか、もしくは、タグリスト取得部111により、図3（A）に示すXML文書から直接的に抽出して作成されるもので、図3（B）に示す例では、図3（A）のXML文書から、ただ単に要素内容「STUV」、「1111」、「A社」、「ABCD」、「980」を取り除いた構成となっている。

【0046】そして、図3（C）には、図3（A）に示すXML文書に対し、構造化文書圧縮部112によるタグ圧縮処理を施した結果、即ち、図3（A）に示すXML文書中のタグを区切りコード「,」に置き換えたXML圧縮文書が示されている。これらの図3（A）～図3（C）を比較対照しても明らかなように、タグリスト中の各タグとXML圧縮文書中の各区切りコード「,」とは一対一で対応するとともに、XML圧縮文書において区切りコード「,」はタグの位置に対応して配置される。また、XML文書中の要素内容は、XML圧縮文書中においてそのまま記述されている。従って、第1実施形態のXML圧縮文書は、XML文書と同様、自由なデータ構造を表現することができるほか、テキストで記述されるため、視認性を維持することもできる。

【0047】一方、図2に示すメモリ展開部210においては、まず、復元・展開対象のXML圧縮文書に対応付けられたタグリストが、ハードディスク400から読出部500により読み出されて解析部211に入力される。この解析部211においては、入力されたタグリストが解析され、その解析結果として、復元・展開対象の複数のXML圧縮文書について共通のデータ構造が得られる。そして、解析部211で得られたデータ構造は、複製部212により、XML圧縮文書毎にメモリ214上に複製データ構造として展開・複製される。

【0048】この後、各XML圧縮文書中の要素内容が、書出部213により、複製データ構造におけるタグの位置と、各XML圧縮文書中の区切りコード「,」の位置との対応をとりながら、メモリ214上における複製データ構造の所定領域に書き出される。これにより、各XML圧縮文書（各構造化文書）が、メモリ214上に展開された状態で復元されることになる。

【0049】このように、本発明の第1実施形態によれば、構造化文書圧縮部112により生成された各XML圧縮文書においては、タグが区切りコード「,」に置換されているだけで、データ内容（要素内容）はそのまま記述されているので、XML文書（構造化文書）の利点であるデータ構造の視認性や柔軟性／拡張性を生かしたまま、XML文書を圧縮してXML文書のデータ量を削減することができる。

【0050】従って、XML文書（XML圧縮文書）を格納するための記憶領域の容量を削減することができ、XML圧縮文書を格納する記憶媒体（本実施形態ではハードディスク400）の記憶領域を有効に利用できるようになるほか、XML文書データの伝送速度を高速化することができる。

【0051】また、複数のXML文書の圧縮結果は、データ構造（一つのタグリスト）とデータ内容（複数のXML圧縮文書）とに分離されているので、第1実施形態のメモリ展開部210では、一つのタグリストに対する解析処理を解析部211において一度だけ行ない、複数のXML圧縮文書に共通のデータ構造を取得してしまえば、後は、取得されたデータ構造を複製部212により複製して流用することで、XML圧縮文書毎に一々タグ解析を行なう必要がなくなる。

【0052】これにより、同一のデータ構造を有する多数のXML文書を取り扱う際に、メモリ展開部210の解析部211において無駄なタグ解析を行なう必要が一切なくなるので、タグ解析の負荷が大幅に低減され、XML文書をメモリ214に展開する際の処理速度を飛躍的に高速化することができる。

【0053】〔2〕第2実施形態の説明

図4は本発明の第2実施形態における構造化文書圧縮装置の構成を示すブロック図、図5は本発明の第2実施形態における構造化文書復元装置の構成を示すブロック図

である。図4に示す構造化文書圧縮装置120および図5に示す構造化文書復元装置220は、XML文書に対する処理を行なうための構造化文書処理システムに含まれて、この構造化文書処理システムの一部を構成するものである。

【0054】この第2実施形態の構造化文書圧縮装置120は、XML文書を圧縮するためのもので、図4に示すように、入力部121、タグ検出部122、タグ圧縮部123および出力部124をそなえて構成されている。

10 なお、構造化文書圧縮装置120は、第1実施形態の構造化文書圧縮部112として用いることも可能である。

【0055】また、第2実施形態の構造化文書復元装置220は、構造化文書圧縮装置120により生成されたXML圧縮文書（圧縮データ）をXML文書に復元するためのもので、図5に示すように、入力部221、タグリスト保持部222、区切りコード検出部223、タグ復元部224および出力部225をそなえて構成されている。

20 【0056】ここで、構造化文書圧縮装置120および構造化文書復元装置220は、同一のコンピュータ上、もしくは、それぞれ異なるコンピュータ上にそなえられている。そして、構造化文書圧縮装置120を成す入力部121、タグ検出部122、タグ圧縮部123および出力部124としての機能は、コンピュータ上で所定のプログラム（構造化文書圧縮プログラム）を実行することにより実現される。同様に、構造化文書復元装置220を成す入力部221、区切りコード検出部223、タグ復元部224および出力部225としての機能は、コンピュータ上で所定のプログラム（構造化文書復元プログラム）を実行することにより実現されるようになって

30 いる。

【0057】さて、図4に示す構造化文書圧縮装置120において、入力部121は、圧縮対象のXML文書を、ハードディスク等（例えば図1の符号300参照）から取り込むものであり、タグ検出部122は、入力部121により取り込まれたXML文書中のタグを検出するものである。

40 【0058】タグ圧縮部123は、タグ検出部122により検出されたタグを、所定の区切りコードに置き換えて圧縮するものである。なお、第2実施形態では、第1実施形態と同様、図6（C）を参照しながら後述するごとく、所定の区切りコードとして「,」を用いる。また、2種類の区切りコード「,」および「/」を準備しておき、タグ圧縮部123が、これら2種類の区切りコードを開始タグと終了タグとで使い分け、図6（D）を参照しながら後述するごとく、開始タグを「,」に置き換えるとともに終了タグを「/」に置き換えるように構成してもよい。出力部124は、タグ圧縮部123を用

50 いて生成されたXML圧縮文書を、圧縮結果として、ハ

ードディスク等（例えば図1、図2、図15～図17の符号400、410、420、440参照）に出力・格納するものである。

【0059】一方、図5に示す構造化文書復元装置220において、入力部221は、復元対象のXML圧縮文書を、記憶媒体等（例えば図1、図2、図15～図17に示すハードディスク400、410、420、440）から取り込むものである。タグリスト保持部222は、XML文書におけるタグを出現順序に従ってリスト化したタグリスト（例えば図6（B）参照）を予め保持するものである。このタグリスト保持部222には、予め生成されたタグリストをデータベース（図示略）等から取得して格納する。

【0060】なお、第2実施形態においても、第1実施形態と同様、処理対象となる複数のXML文書が同一のデータ構造を有していることを前提としており、タグリストは、第1実施形態において前述した通り、これら複数のXML文書により共有され、各XML文書において、どのようなタグがどのような順序で出現するかを示すものである。

【0061】区切りコード検出部223は、入力部221により取り込まれたXML圧縮文書中の区切りコードを検出するものである。タグ復元部224は、タグリスト保持部222に保持されたタグリストにおけるタグの位置と、区切りコード検出部223により検出された区切りコードの位置とを対応させながら、その区切りコードを、タグリストにおける対応するタグに置き換えて復元するものである。出力部225は、タグ復元部224を用いて復元されたXML文書を、復元結果として、記憶媒体等（例えば図1に示すハードディスク300）に出力・格納するものである。

【0062】次に、上述のごとく構成された、第2実施形態の構造化文書圧縮装置120および構造化文書復元装置220の動作について説明する。図4に示す構造化文書圧縮装置120においては、まず、圧縮対象のXML文書を入力部121により取り込み、そのXML文書中のタグをタグ検出部122により探索する。タグ以外の部分（つまり要素内容の部分）はそのまま出力部124へ送られるが、タグ検出部122によりタグが検出されると、そのタグは、タグ圧縮部123により所定の区切りコードに置き換えられて圧縮されてから、出力部124へ送られる。このような単純な置換処理によって生成されたXML圧縮文書が、圧縮結果として出力部124から出力される。

【0063】ここで、図6（A）～図6（D）を参照しながら第2実施形態における具体的なXML文書の圧縮状態について説明する。なお、図6（A）～図6（D）はいずれも第2実施形態におけるデータ例を示すもので、図6（A）はXML文書の一例を示す図、図6

（B）は図6（A）に示すXML文書に対応するタグリ

ストを示す図、図6（C）は図6（A）に示すXML文書の圧縮状態の一例を示す図、図6（D）は図6（A）に示すXML文書の圧縮状態の他例を示す図である。

【0064】図6（A）には、圧縮前つまり圧縮対象のXML文書の一例として、発注伝票をXMLにより記述した例が示されている。特に、図6（A）では、図3（A）を参照しながら前述した発注伝票の一部分（商品のメーカ、製品番号および価格にかかる記述部分）が抽出されて示されている。

【0065】また、図6（B）は、図6（A）に示したXML文書のタグリストを示しており、このようなタグリストが、予め何らかの手段により抽出・生成されて、構造化文書復元装置220のタグリスト保持部222に格納されている。なお、図6（B）に示すタグリストでは、タグの前後に付される括弧表示（“<”および“>”）が省略されている。

【0066】そして、図6（C）には、図6（A）に示すXML文書に対し、タグ圧縮部123によるタグ圧縮処理を施した結果、即ち、図6（A）に示すXML文書中のタグを区切りコード「,」に置き換えたXML圧縮文書が示されている。また、図6（D）には、同一のXML文書についての他の圧縮結果が示されている。つまり、図6（D）に示すXML圧縮文書は、タグ圧縮部123によるタグ圧縮処理に際して、XML文書中の開始タグを「,」に置き換え、XML文書中の終了タグを「/」に置き換えたものである。

【0067】これらの図6（A）～図6（D）を比較対照しても明らかなように、タグリスト中の各タグとXML圧縮文書中の各区切りコード「,」または「/」とは一対一で対応するとともに、XML圧縮文書において区切りコード「,」または「/」はタグの位置に対応して配置される。また、XML文書中の要素内容は、XML圧縮文書中においてそのまま記述されている。

【0068】従って、第2実施形態のXML圧縮文書によっても、XML文書と同様の自由なデータ構造表現が可能であり、要素内容の視認性が維持される。特に、図6（D）に示すXML圧縮文書では、2種類の区切りコード「,」と「/」とがそれぞれ開始タグと終了タグとに対応して用いられるので、開始タグおよび終了タグの位置を視認することも可能になる。

【0069】一方、図5に示す構造化文書復元装置220においては、まず、例えば図6（C）もしくは図6（D）に示すような復元対象のXML圧縮文書を入力部221により取り込み、そのXML圧縮文書中の区切りコード（「,」もしくは「,」と「/」）を区切りコード検出部223により探索する。

【0070】区切りコード以外の部分（つまり要素内容の部分）は、そのまま出力部225へ送られるが、区切りコード検出部223により区切りコードが検出されると、その区切りコードは、タグ復元部224により、そ

の区切りコードに対応したタグに置き換えられてから、出力部 225 へ送られる。このような単純な置換処理によって例えば図 6 (A) に示すような XML 文書が復元され、出力部 225 から出力される。

【0071】タグ復元部 224 による置換処理に際しては、タグリスト保持部 222 に保持されたタグリストにおけるタグの位置と、区切りコード検出部 223 により検出された区切りコードの位置との対応をとることにより、検出された区切りコードがタグリスト中のどのタグに対応するかが認識されるので、区切りコードを適切なタグ（対応するタグ）に復元することができる。このようにして、XML 圧縮文書は、元の XML 文書に復元される。

【0072】このように、本発明の第 2 実施形態の構造化文書圧縮装置 120 によれば、XML 文書中で検出されたタグを所定の区切りコードに置換するという極めて単純な置換処理によって、XML 文書（構造化文書）の利点であるデータ構造の視認性や柔軟性／拡張性の高さを生かしたまま、XML 文書を圧縮して XML 文書のデータ量を削減することができる。

【0073】従って、第 2 実施形態においても、第 1 実施形態と同様、XML 文書（XML 圧縮文書）を格納するための記憶領域の容量を削減することができ、XML 圧縮文書を格納する記憶媒体（例えば図 1、図 2、図 15～図 17 に示すハードディスク 400、410、420、440）の記憶領域を有効に利用できるようになるほか、XML 文書データの伝送速度を高速化することができる。

【0074】また、第 2 実施形態の構造化文書復元装置 220 によれば、XML 圧縮文書中で検出された区切りコードを、その XML 圧縮文書についてのタグリスト中のタグと対応させながら、所定のタグに置き換えるという簡易な置換処理によって、XML 圧縮文書を極めて容易に元の XML 文書に復元することができるという利点もある。

【0075】〔3〕第 3 実施形態の説明

図 7 は本発明の第 3 実施形態における構造化文書圧縮装置の構成を示すブロック図、図 8 は本発明の第 3 実施形態における構造化文書復元装置の構成を示すブロック図である。なお、図中、既述の符号と同一の符号は同一の部分もしくはほぼ同一の部分を示しているので、その詳細な説明は省略する。

【0076】図 7 に示す構造化文書圧縮装置 130 および図 8 に示す構造化文書復元装置 230 も、第 2 実施形態と同様、XML 文書に対する処理を行なうための構造化文書処理システムに含まれて、この構造化文書処理システムの一部を構成するもので、それぞれ、図 4 に示す構造化文書圧縮装置 120 および図 5 に示す構造化文書復元装置 220 とほぼ同様に構成されている。

【0077】ただし、第 3 実施形態の構造化文書圧縮装

置 130 は、XML 文書のタグ内に属性が記述されている場合には、その属性を圧縮後も残すことにより属性の視認性をも維持しながら、XML 文書の圧縮を行なえるように構成したもので、図 7 に示すように、第 2 実施形態と同様の入力部 121、タグ検出部 122、タグ圧縮部 123 および出力部 124 のほか、さらに、属性付きタグ検出部 131 および属性付きタグ圧縮部 132 をそなえて構成されている。なお、この構造化文書圧縮装置 130 も、第 1 実施形態の構造化文書圧縮部 112 として用いることが可能である。

【0078】また、第 3 実施形態の構造化文書復元装置 230 は、構造化文書圧縮装置 130 により生成された XML 圧縮文書（圧縮データ）を XML 文書に復元するためのもので、図 8 に示すように、第 2 実施形態と同様の入力部 221、タグリスト保持部 222、区切りコード検出部 223、タグ復元部 224 および出力部 225 のほか、さらに、属性リスト保持部 231、属性付きタグ検出部 232 および属性付きタグ復元部 233 をそなえて構成されている。

【0079】ここで、第 3 実施形態の構造化文書圧縮装置 130 および構造化文書復元装置 230 も、同一のコンピュータ上、もしくは、それぞれ異なるコンピュータ上にそなえられている。そして、構造化文書圧縮装置 130 を成す入力部 121、タグ検出部 122、タグ圧縮部 123、出力部 124、属性付きタグ検出部 131 および属性付きタグ圧縮部 132 としての機能は、コンピュータ上で所定のプログラム（構造化文書圧縮プログラム）を実行することにより実現される。同様に、構造化文書復元装置 230 を成す入力部 221、区切りコード検出部 223、タグ復元部 224、出力部 225、属性付きタグ検出部 232 および属性付きタグ復元部 233 としての機能は、コンピュータ上で所定のプログラム（構造化文書復元プログラム）を実行することにより実現されるようになっている。

【0080】さて、図 7 に示す構造化文書圧縮装置 130 において、属性付きタグ検出部 131 は、タグ検出部 122 に含まれており、タグ検出部 122 により検出されたタグが属性値をもつ属性付きタグであるか否かを検出するものである。なお、属性付きタグとは、そのタグ内に、要素内容に付加したい情報（属性）を記述されたものである。その属性は、具体的には図 9 (A) を参照しながら後述するごとく、開始タグ内において、要素名の後にスペースを空け「属性名＝「属性値」」として記述される。つまり、属性付きタグは、一般的には「<要素名 属性名＝「属性値」>」と記述される。

【0081】属性付きタグ圧縮部 132 は、属性付きタグ検出部 131 により検出された属性付きタグを、そのタグ内に記述された属性値と所定の区切りコードとにより置き換えて圧縮するものである。この属性付きタグ圧縮部 132 によって置き換えられる区切りコードとして

は、例えば図9 (C) や図9 (D) を参照しながら後述するごとく「, 」あるいは「=」を用いる。

【0082】また、本実施形態では、属性付きタグ圧縮部132により属性付きタグを属性値と区切りコードとに置き換える際、その区切りコードは属性値の前後に配置されるようになっている〔図9 (C) や図9 (D) 参照〕。例えば「<要素名 属性名= '属性値' >」は「, 属性値, 」もしくは「, 属性値=」という圧縮記述に置き換えられる。また、複数の属性をもつタグ、例えば「<要素名 属性名1= '属性値1' 属性名2=

'属性値2' >」と記述されたタグは「, 属性値1, 属性値2, 」もしくは「, 属性値1=属性値2=」という圧縮記述に置き換えられる。

【0083】なお、第3実施形態の出力部124は、タグ圧縮部123および属性付きタグ圧縮部132を用いて生成されたXML圧縮文書を、圧縮結果として、ハードディスク等（例えば図1, 図2, 図15～図17の符号400, 410, 420, 440参照）に出力・格納するようになっている。

【0084】一方、図8に示す構造化文書復元装置230において、属性リスト保持部231は、XML圧縮文書における属性名を出現順序に従ってリスト化した属性リストを予め保持するものである。この属性リスト保持部231には、予め生成された属性リストをデータベース（図示略）等から取得して格納する。

【0085】ここで、第3実施形態の属性リストは、実際には、図9 (B) を参照しながら後述するごとく、タグリストに含まれる形で与えられるものである。このため、図8では、属性リスト保持部231がタグリスト保持部222に含まれている。以下では、属性リストを含むタグリストのことをタグ・属性リストと表記する場合がある。このようなタグ・属性リストにおいては、そのリストに記入された文字列が属性名である場合、そのことが明確に分かるように、例えば図9 (B) に示すごとく、その文字列の前（左側）に、例えばコード「=」を付与している。

【0086】また、第3実施形態においても、第1実施形態と同様、処理対象となる複数のXML文書が同一のデータ構造を有していることを前提としており、タグリストおよび属性リスト（タグ・属性リスト）は、これら複数のXML文書により共有され、各XML文書において、どのようなタグがどのような順序で出現するか、あるいは、どのような属性がどのような順序で出現するかを示すものである。

【0087】属性付きタグ検出部232は、タグ復元部224に含まれており、タグ復元部224で復元対象となったタグが属性付きタグに復元されるべきものであるか否かを検出するものである。このとき、属性付きタグ検出部232は、区切りコードの配置状態や区切りコードの種類を認識することにより、もしくは、XML圧縮

文書内の属性値とタグ・属性リスト内の属性名との対応関係を参照・認識することにより、復元対象のタグが、属性付きタグに復元されるべきもの、即ち、属性をもつものであるか否かを検出している。

【0088】属性付きタグ復元部233は、属性付きタグ検出部232により復元対象として検出されたタグを、そのタグに対応した属性を有する属性付きタグに復元するものである。第3実施形態においては、復元対象となるXML圧縮文書のうち属性付きタグに対応する部分は、まず、タグ復元部224において要素名のみを含む通常のタグ「<要素名>」として復元される。第3実施形態の属性付きタグ復元部233は、属性付きタグについての属性値と属性リストにおける属性名とを対応させて、属性付きタグ内の属性を復元するものである。

【0089】より具体的に説明すると、属性付きタグ復元部233は、復元すべき属性に対応する属性名を属性リスト（タグ・属性リスト）から読み出し、復元すべき属性に相当する区切りコードとこの区切りコードに組み合わせられたデータ（属性値）とを通常の属性記述に置き換えることで、タグ復元部224で復元されたタグ内に属性を復元させ、属性付きタグの復元を行なうようになっている。例えば「属性値, 」または「属性値=」という属性の圧縮記述は「属性名= '属性値' 」に置き換えられ、「属性値1, 属性値2, 」または「属性値1=属性値2=」という属性の圧縮記述は「属性名1= '属性値1' 属性名2= '属性値2' 」に置き換えられる。

【0090】なお、第3実施形態の出力部225は、タグ復元部224および属性付きタグ復元部233を用いて復元されたXML文書を、復元結果として、記憶媒体等（例えば図1に示すハードディスク300）に出力・格納するようになっている。次に、上述のごとく構成された、第3実施形態の構造化文書圧縮装置130および構造化文書復元装置230の動作について説明する。

【0091】図7に示す構造化文書圧縮装置130においては、まず、圧縮対象のXML文書を入力部121により取り込み、そのXML文書中のタグをタグ検出部122により探索する。タグ以外の部分（つまり要素内容の部分）は、そのまま出力部124へ送られるが、タグ検出部122によりタグが検出されると、属性付きタグ検出部131により、そのタグが属性付きタグであるか否かが検出される。

【0092】属性付きタグでない場合、第2実施形態で説明した通り、そのタグは、タグ圧縮部123により所定の区切りコードに置き換えられて圧縮されてから、出力部124へ送られる。一方、属性付きタグである場合、そのタグは、属性付きタグ圧縮部132により、そのタグ内に記述された属性値と所定の区切りコードとに置き換えられて圧縮されてから、出力部124へ送られる。

【0093】第3実施形態においては、XML文書が属

性付きタグを有している場合、上述のような単純な置換処理により、要素内容とともに属性値を残したままのXML圧縮文書が生成され圧縮結果として出力部124から出力される。ここで、図9(A)～図9(D)を参照しながら、第3実施形態における具体的なXML文書の圧縮状態について説明する。なお、図9(A)～図9

(D)はいずれも第3実施形態におけるデータ例を示すもので、図9(A)はXML文書の一例を示す図、図9(B)は図9(A)に示すXML文書に対応するタグ・属性リストを示す図、図9(C)は図9(A)に示すXML文書の圧縮状態の一例を示す図、図9(D)は図9(A)に示すXML文書の圧縮状態の他例を示す図である。

【0094】図9(A)には、圧縮前つまり圧縮対象のXML文書の一例として、発注伝票をXMLにより記述した例が示されている。特に、図9(A)では、図6(A)に示した例とほぼ同様の発注伝票の一部分が抽出されて示されている。この図9(A)に示す例では、さらに、製品番号を要素名としてもつ開始タグが属性を有している。即ち、その開始タグ(属性付きタグ)内には、属性として「製品名=「ABCD」色=「青」」が記述されている。ここで、「製品名」および「色」が属性名であり、「ABCD」および「青」が属性値である。

【0095】また、図9(B)は、図9(A)に示したXML文書のタグ・属性リストを示しており、このようなタグ・属性リストが、予め何らかの手段により抽出・生成されて、構造化文書復元装置230のタグリスト保持部222および属性リスト保持部231に格納されている。この図9(B)に示すタグ・属性リストは、図6(B)に示したタグリストに、製品番号の属性名に係る項目、つまり「=製品名」および「=色」をさらに追加したものである。

【0096】そして、図9(C)には、図9(A)に示すXML文書に対し、タグ圧縮部123および属性付きタグ圧縮部132による圧縮処理を施した結果、即ち、図9(A)に示すXML文書中のタグを区切りコード「,」に置き換えるとともに、属性を「属性値+区切りコード「,」」に置き換えたXML圧縮文書が示されている。つまり、図9(A)における属性付きタグ「<製品番号 製品名=「ABCD」色=「青」>」は、図9(C)に示すXML圧縮文書では、「, ABCD, 青,」に置き換えられている。

【0097】また、図9(D)には、同一のXML文書についての他の圧縮結果が示されている。つまり、図9(D)に示すXML圧縮文書は、タグ圧縮部123によるタグ圧縮処理に際して、XML文書中の開始タグを「,」に置き換え、XML文書中の終了タグを「/」に置き換えるとともに、属性付きタグ圧縮部132による圧縮処理に際し、属性値に付加する区切りコードとし

て「=」を用いたものである。従って、図9(A)における属性付きタグ「<製品番号 製品名=「ABCD」色=「青」>」は、図9(D)に示すXML圧縮文書では、「, ABCD=青=」に置き換えられている。

【0098】これらの図9(A)～図9(D)を比較対照しても明かなように、タグ・属性リスト中の各タグとXML圧縮文書中の各区切りコード「,」または「/」とは一対一で対応するとともに、XML圧縮文書において区切りコード「,」または「/」はタグの位置に対応して配置される。また、XML文書中の要素内容は、XML圧縮文書中においてそのまま記述されている。さらに、XML文書中の属性値は、XML圧縮文書中において、右側に区切りコード「,」または「=」を付加された状態で、そのまま記述されている。

【0099】従って、第3実施形態のXML圧縮文書によっても、XML文書と同様の自由なデータ構造表現が可能であり、要素内容のみならず属性値についても視認性が維持される。特に、図9(D)に示すXML圧縮文書では、3種類の区切りコード「,」と「/」と「=」がそれぞれ開始タグと終了タグと属性とに対応して用いられるので、開始タグ、終了タグおよび属性(属性付きタグ)の位置を視認することも可能になる。

【0100】一方、図8に示す構造化文書復元装置230においては、まず、例えば図9(C)もしくは図9(D)に示すような復元対象のXML圧縮文書を入力部221により取り込み、そのXML圧縮文書中の区切りコード(「,」や「/」)を区切りコード検出部223により探索する。

【0101】区切りコードおよび属性値以外の部分(つまり要素内容の部分)は、そのまま出力部225へ送られるが、区切りコード検出部223により区切りコードが検出されると、その区切りコードは、タグ復元部224により、その区切りコードに対応したタグに置き換えられる。タグ復元部224による置換処理に際しては、第2実施形態と同様、タグリスト保持部222に保持されたタグリストにおけるタグの位置と、区切りコード検出部223により検出された区切りコードの位置との対応をとることにより、検出された区切りコードがタグリスト中のどのタグに対応するかが認識されるので、区切りコードを適切なタグ(対応するタグ)に復元することができる。

【0102】そして、第3実施形態では、属性付きタグ検出部232により、タグ復元部224で復元対象となったタグが属性をもつものであるか否かを検出し、属性をもたないものであると認識された場合には、タグ復元部224で復元されたタグ(属性をもたないタグ)は、そのまま出力部225へ送られる。一方、属性をもつものであると認識された場合には、タグ復元部224により要素名のみを含む状態で復元された通常のタグ(例えば製品番号)内に、そのタグに対応する属性を、属

性付きタグ復元部 233 によって復元させてから、出力部 225 へ送られる。

【0103】例えば図 9 (C) や図 9 (D) に示す圧縮記述「, ABCD, 青, 」や「, ABCD=青=」については、その圧縮記述の最初の区切りコード「, 」が検出され、その区切りコードが「製品番号」に対応するものであることが認識される。さらに、図 9 (B) に示すタグ・属性リストを参照することにより、上記圧縮記述に対応するタグは、「製品名」および「色」を属性名とする 2 つの属性をもつことが認識される。このような属性情報の認識に応じて、属性付きタグ復元部 233 により、上記圧縮記述は、図 9 (A) に示すような属性付きタグ「<製品番号製品名= 'ABCD' 色= '青' >」に変換・復元される。

【0104】このように、本発明の第 3 実施形態の構造化文書圧縮装置 130 によれば、第 2 実施形態の構造化文書圧縮装置 120 と同様の作用効果が得られるほか、タグが属性値をもつ属性付きタグである場合には、その属性付きタグが属性値および所定の区切りコードに置き換えられて圧縮される。これにより、XML 圧縮文書において属性値がそのまま記述されるので、要素内容だけでなく属性値の視認性を保ちながら XML 文書の圧縮を行なうことができる。

【0105】また、第 3 実施形態の構造化文書復元装置 230 によれば、第 2 実施形態の構造化文書圧縮装置 220 と同様の作用効果が得られるほか、上述のような圧縮を施された属性付きタグが復元対象になると、その属性付きタグについての属性値と XML 圧縮文書についてのタグ・属性リスト中の属性名とを対応させることにより、属性付きタグを極めて容易に復元することができる。

【0106】〔4〕第 4 実施形態の説明

図 10 は本発明の第 4 実施形態における構造化文書圧縮装置の要部構成を示すブロック図である。この図 10 に示す構造化文書圧縮装置 140 は、図 4 に示す構造化文書圧縮装置 120 の前段に、さらに、入力部 141、タグリスト保持部 142、タグ並び替え部 143 および省略タグ補完部 144 をそなえて構成されたものである。なお、この構造化文書圧縮装置 140 も、第 1 実施形態の構造化文書圧縮部 112 として用いることが可能である。また、構造化文書圧縮装置 140 の要部を成す入力部 141、タグ並び替え部 143 および省略タグ補完部 144 も、コンピュータ上で所定のプログラム（構造化文書圧縮プログラム）を実行することにより実現される。

【0107】さて、図 10 に示す構造化文書圧縮装置 140 において、入力部 141 は、圧縮対象の XML 文書を、ハードディスク等（例えば図 1 の符号 300 参照）から取り込むものである。また、タグリスト保持部 142 は、所定のデータ構造を定義すべく所定の順序でタグ

を並べたタグリストを予め保持するものである。より詳細に説明すると、第 3 実施形態においても、第 1 および第 2 実施形態と同様、処理対象となる複数の XML 文書が、同一のデータ構造を有していることを前提としている。そして、タグリスト保持部 142 に保持されるタグリストは、構造化文書圧縮装置 220 のタグリスト保持部 222 に保持されるタグリストと同様、これら複数の XML 文書により共有され、各 XML 文書において、どのようなタグがどのような順序で出現するかを示すものである。なお、タグリスト保持部 142 には、圧縮処理対象となる XML 文書について予め生成されたタグリストが、データベース（図示略）等から取得して格納される。

【0108】タグ並び替え部 143 は、入力された XML 文書とタグリストとを比較し、XML 文書におけるタグの記述順序をタグリストにおけるタグの配列順序（所定の順序）に合わせるように、圧縮前の XML 文書のタグを並び替えるものである。このとき、対になる開始タグと終了タグとの順序を変更する場合、タグ並び替え部 143 は、これらの開始タグと終了タグとの間に記述された要素内容も一緒に移動させる。

【0109】省略タグ補完部 144 は、タグリスト保持部 142 に保持されたタグリストに従って、タグ並び替え部 143 による処理を施された XML 文書中で省略されているタグを補完するものである。つまり、省略タグ補完部 144 は、入力された XML 文書とタグリストとを比較し、その XML 文書中から欠落しているタグを検出すると、欠落タグに対応するタグをタグリストから読み出し、その欠落タグを補完するものである。このとき、対になる開始タグと終了タグとを補完する場合、省略タグ補完部 144 は、これらの開始タグと終了タグとの間に記述されるべき要素内容を空のままとする。

【0110】そして、タグ並び替え部 143 および省略タグ補完部 144 による処理を施された XML 文書は、第 2 実施形態の構造化文書圧縮装置 120 に入力されるようになっている。

【0111】次に、上述のごとく構成された、第 4 実施形態の構造化文書圧縮装置 140 の動作について説明する。図 10 に示す構造化文書圧縮装置 140 においては、まず、圧縮対象の XML 文書を入力部 141 により取り込み、タグ並び替え部 143 において、その XML 文書とタグリストとが比較され、万一、XML 文書中にタグの記述順序の逆転等の不備がある場合には、XML 文書におけるタグの記述順序がタグリストにおけるタグの配列順序に合うように圧縮前の XML 文書のタグが並び替えられる。

【0112】そして、並び替え処理を施された XML 文書は、省略タグ補完部 144 に入力され、この省略タグ補完部 144 において、その XML 文書とタグリストとが比較され、その XML 文書中から欠落しているタグが

検出されると、欠落タグに対応するタグがタグリストから読み出され、その欠落タグが補完される。

【0113】ここで、図11(A)～図11(C)を参照しながら、第4実施形態における具体的なXML文書の圧縮状態について説明する。なお、図11(A)～図11(C)はいずれも第4実施形態におけるデータ例を示すもので、図11(A)はタグリストの一例を示す図、図11(B)はタグの記述に不備のあるXML文書の一例を示す図、図11(C)は図11(B)に示すXML文書を図11(A)に示すタグリストに従って正規化した結果を示す図である。

【0114】図11(A)には、図6(B)に示したものと全く同じタグリストが示されており、ここでは、この図11(A)に示すタグリストに従って、タグ並び替え部143および省略タグ補完部144による処理をXML文書に施す場合について説明する。その処理対象となるXML文書は、例えば図11(B)に示すものである。

【0115】まず、タグ並び替え部143において、図11(A)のタグリストと図11(B)のXML文書とを比較することにより、図11(B)のXML文書では、価格についてのタグおよび要素内容「300」と、製品番号についてのタグおよび要素内容「B7」との配置順序が逆転していることが認識され、その順序が並び替えられる。

【0116】そして、省略タグ補完部144において、上述のごとく順序を並び替えられたXML文書と図11(A)のタグリストとを比較することにより、そのXML文書では、メーカーについてのタグが欠落していることが認識され、メーカーについての開始タグと終了タグとが空要素の状態では補完される。その結果、図11(B)に示すようにタグの記述に不備のあったXML文書が、図11(B)に示すタグリストに応じたデータ構造をもつXML文書に修正され、図11(C)に示すようなXML文書に正規化(整頓)される。

【0117】つまり、構造化文書圧縮装置120による圧縮対象である、全てのXML文書に対し、タグ並び替え部143および省略タグ補完部144による処理を施すことによって、全てのXML文書が、タグリストに応じたデータ構造をもつXML文書となるように正規化される。

【0118】そして、上述のごとく正規化されたXML文書が、構造化文書圧縮装置120に入力され、第2実施形態で前述したように圧縮される。なお、当然、不備のないXML文書は、タグ並び替え部143および省略タグ補完部144をそのまま通過して、構造化文書圧縮装置120に入力される。

【0119】ところで、図12は本発明の第4実施形態における構造化文書圧縮装置の変形例の要部構成を示すブロック図である。この図12に示す構造化文書圧縮装

置150は、図7に示す構造化文書圧縮装置130の前に、さらに、入力部151、タグ・属性リスト保持部152、タグ・属性並び替え部153および省略タグ・属性補完部154をそなえて構成されたものである。なお、この構造化文書圧縮装置150も、第1実施形態の構造化文書圧縮部112として用いることが可能である。また、構造化文書圧縮装置150の要部を成す入力部151、タグ・属性並び替え部153および省略タグ・属性補完部154も、コンピュータ上で所定のプログラム(構造化文書圧縮プログラム)を実行することにより実現される。

【0120】さて、図12に示す構造化文書圧縮装置150において、入力部151は、属性付きタグを含む圧縮対象のXML文書を、ハードディスク等(例えば図1の符号300参照)から取り込むものである。また、タグ・属性リスト保持部152は、所定のデータ構造を定義すべく所定の順序で並べたタグと属性名とをもつタグ・属性リストを予め保持するものである。より詳細に説明すると、この第4実施形態の変形例においても、第1～第3実施形態と同様、処理対象となる複数のXML文書が、同一のデータ構造を有していることを前提としている。そして、タグ・属性リスト保持部152に保持されるタグ・属性リストは、構造化文書圧縮装置230のタグ・属性リストと同様、これら複数のXML文書により共有され、各XML文書において、どのようなタグがどのような順序で出現するか、あるいは、どのような属性がどのような順序で出現するかを示すものである。なお、タグ・属性リスト保持部152には、圧縮処理対象となるXML文書について予め生成されたタグ・属性リストが、データベース(図示略)等から取得して格納される。

【0121】タグ・属性並び替え部153は、入力されたXML文書とタグ・属性リストとを比較し、XML文書におけるタグおよび属性の記述順序をタグ・属性リストにおけるタグおよび属性の配列順序(所定の順序)に合わせるように、圧縮前のXML文書のタグや属性を並び替えるものである。このとき、対になる開始タグと終了タグとの順序を変更する場合、タグ・属性並び替え部153は、これらの開始タグと終了タグとの間に記述された要素内容も一緒に移動させる。

【0122】省略タグ・属性補完部154は、タグ・属性リスト保持部152に保持されたタグ・属性リストに従って、タグ・属性並び替え部153による処理を施されたXML文書中で省略されているタグや属性を補完するものである。つまり、省略タグ・属性補完部154は、入力されたXML文書とタグリストとを比較して、そのXML文書中から欠落しているタグや属性を検出すると、欠落タグや欠落属性に対応するタグあるいは属性名をタグ・属性リストから読み出し、その欠落タグや欠落属性を補完するものである。このとき、対になる開始

タグと終了タグとを補完する場合、省略タグ・属性補完部 154 は、これらの開始タグと終了タグとの間に記述されるべき要素内容を空のままとする。また、属性を補完する場合、省略タグ・属性補完部 154 は、属性値としてデフォルト値等を設定する。

【0123】そして、タグ・属性並び替え部 153 および省略タグ・属性補完部 154 による処理を施された XML 文書は、第 3 実施形態の構造化文書圧縮装置 130 に入力されるようになっている。

【0124】次に、上述のごとく構成された、第 4 実施形態の変形例の構造化文書圧縮装置 150 の動作について説明する。図 12 に示す構造化文書圧縮装置 150 においては、まず、圧縮対象の XML 文書を入力部 151 により取り込み、タグ・属性並び替え部 153 において、その XML 文書とタグ・属性リストとが比較され、万一、XML 文書中にタグや属性の記述順序の逆転等の不備がある場合には、XML 文書におけるタグや属性の記述順序がタグ・属性リストにおけるタグや属性の配列順序に合うように、圧縮前の XML 文書のタグや属性が並び替えられる。

【0125】そして、並び替え処理を施された XML 文書は、省略タグ・属性補完部 154 に入力され、この省略タグ・属性補完部 154 において、その XML 文書とタグ・属性リストとが比較され、その XML 文書中から欠落しているタグや属性が検出されると、欠落タグや欠落属性に対応するタグや属性がタグ・属性リストから読み出され、その欠落タグや欠落属性が補完される。

【0126】その結果、タグや属性の記述に不備のあった XML 文書が、タグ・属性リストに応じたデータ構造をもつ XML 文書に修正され正規化（整頓）される。つまり、構造化文書圧縮装置 130 による圧縮対象である、全ての XML 文書に対し、タグ・属性並び替え部 153 および省略タグ・属性補完部 154 による処理を施すことによって、全ての XML 文書が、タグ・属性リストに応じたデータ構造をもつ XML 文書となるように正規化される。

【0127】そして、上述のごとく正規化された XML 文書が、構造化文書圧縮装置 130 に入力され、第 3 実施形態で前述したように圧縮される。なお、当然、不備のない XML 文書は、タグ・属性並び替え部 153 および省略タグ・属性補完部 154 をそのまま通過して、構造化文書圧縮装置 130 に入力される。

【0128】このように、本発明の第 4 実施形態における構造化文書圧縮装置 140、150 によれば、所定のデータ構造を定義する、タグリストまたはタグ・属性リストに従って、圧縮前の XML 文書のタグや属性が所定の順序に並び替えられるとともに、XML 文書中で省略されているタグや属性が補完される。これにより、タグまたは属性の記述順序の逆転や、タグまたは属性の記述の欠落といった不備をもつ XML 文書は、所定のデータ構

造を有するように正規化される。

【0129】従って、同一のデータ構造を有する多数の XML 文書を圧縮処理対象とする場合、上述のような不備をもつ XML 文書が含まれていても、圧縮処理前に、圧縮処理対象の全ての XML 文書が、タグリストもしくはタグ・属性リストで定義された所定のデータ構造を有するように正規化される。これにより、多数の XML 圧縮文書を、一つのタグリストまたはタグ・属性リストによって確実に管理し取り扱うことが可能になる。

10 【0130】【5】第 5 実施形態の説明

図 13 は本発明の第 5 実施形態における構造化文書圧縮装置および構造化文書復元装置の構成を示すブロック図である。図 13 に示す構造化文書圧縮装置 160 および構造化文書復元装置 260 は、XML 文書に対する処理を行なうための構造化文書処理システムに含まれて、この構造化文書処理システムの一部を構成するものである。

【0131】この第 5 実施形態の構造化文書圧縮装置 160 は、XML 文書を圧縮するためのもので、図 13 に示すように、上述した構造化文書圧縮装置 120、130、140、150 のいずれか一つに対し、さらに、入力部 161、サブ文書抽出部 162、サブ文書統合部 163 および出力部 164 をそなえて構成されている。

【0132】また、第 5 実施形態の構造化文書復元装置 260 は、構造化文書圧縮装置 160 により生成された XML 圧縮文書（圧縮データ）を XML 文書に復元するためのもので、図 13 に示すように、上述した構造化文書復元装置 220、230 のいずれか一方に対し、さらに、入力部 161、サブ文書抽出部 162、サブ文書統合部 163 および出力部 164 をそなえて構成されている。

【0133】ここで、構造化文書圧縮装置 160 および構造化文書復元装置 260 は、同一のコンピュータ上、もしくは、それぞれ異なるコンピュータ上にそなえられている。そして、構造化文書圧縮装置 160 における各種機能は、コンピュータ上で所定のプログラム（構造化文書圧縮プログラム）を実行することにより実現される。同様に、構造化文書復元装置 260 における各種機能は、コンピュータ上で所定のプログラム（構造化文書復元プログラム）を実行することにより実現されるようになっている。

【0134】なお、図 13 に示すように、構造化文書圧縮装置 160 と構造化文書復元部 260 とは、サブ文書抽出部 162 とサブ文書統合部 163 との間に配置される装置が異なる以外は、全く同じ構成になっている。さて、図 13 に示す構造化文書圧縮装置 160 において、入力部 161 は、圧縮対象の XML 文書を、ハードディスク等（例えば図 1 の符号 300 参照）から取り込むものであり、サブ文書抽出部 162 は、入力された XML 文書から、所定の要素名をもつ開始タグと終了タグとで

囲まれた領域をサブ文書として抽出し、そのサブ文書を構造化文書圧縮装置 120, 130, 140, 150 のいずれか一つ（以下、第 5 実施形態では、構造化文書圧縮装置 120 とする）に出力するものである。

【0135】そして、サブ文書統合部 163 は、構造化文書圧縮装置 120 からサブ文書の圧縮結果を受け、その圧縮結果とサブ文書以外の部分とを統合するものであり、出力部 164 は、サブ文書統合部 163 によって統合された XML 圧縮文書を、圧縮結果として、ハードディスク等（例えば図 1, 図 2, 図 15～図 17 の符号 400, 410, 420, 440 参照）に出力・格納するものである。

【0136】一方、図 13 に示す構造化文書復元装置 260 において、入力部 161 は、復元対象の XML 圧縮文書を、記憶媒体等（例えば図 1, 図 2, 図 15～図 17 に示すハードディスク 400, 410, 420, 440）から取り込むものであり、サブ文書抽出部 162 は、入力された XML 圧縮文書から、所定の要素名をもつ開始タグと終了タグとで囲まれた領域をサブ文書として抽出し、そのサブ文書を構造化文書復元装置 220 もしくは 230（以下、第 5 実施形態では、構造化文書復元装置 220 とする）に出力するものである。

【0137】そして、サブ文書統合部 163 は、構造化文書復元装置 220 からサブ文書の復元結果を受け、その復元結果とサブ文書以外の部分とを統合するものであり、出力部 164 は、サブ文書統合部 163 によって統合された XML 文書を、復元結果として、記憶媒体等（例えば図 1 に示すハードディスク 300）に出力・格納するものである。

【0138】なお、構造化文書復元装置 260 において構造化文書復元装置 220 を用いる場合、その構造化文書復元装置 220 において用いられるタグリストは、サブ文書におけるタグを出現順序に従ってリスト化して予め生成されたもので、データベース（図示略）等から取得される。また、構造化文書復元装置 260 において構造化文書復元装置 230 を用いる場合、その構造化文書復元装置 230 において用いられるタグ・属性リストは、サブ文書におけるタグや属性を出現順序に従ってリスト化して予め生成されたもので、やはり、データベース（図示略）等から取得される。

【0139】次に、上述のごとく構成された、第 5 実施形態の構造化文書圧縮装置 160 および構造化文書復元装置 260 の動作について説明する。図 13 に示す構造化文書圧縮装置 160 においては、まず、圧縮対象の XML 文書を入力部 161 により取り込み、サブ文書抽出部 162 により、その XML 文書から、所定の要素名をもつ開始タグと終了タグとで囲まれた領域がサブ文書として抽出され、そのサブ文書が構造化文書圧縮装置 120 に出力される。

【0140】そして、構造化文書圧縮装置 120 におい

ては、入力されたサブ文書に対し、第 2 実施形態で前述した圧縮処理が施され、タグを所定の区切りコードに置き換えるようにして圧縮されたサブ文書が生成される。圧縮されたサブ文書は、サブ文書統合部 163 によりサブ文書以外の部分と統合され、統合された XML 文書が圧縮結果として出力部 164 から出力される。

【0141】ここで、図 14 (A) および図 14 (B) を参照しながら、第 5 実施形態における具体的な XML 文書の圧縮状態について説明する。なお、図 14 (A) および図 14 (B) はいずれも第 5 実施形態におけるデータ例を示すもので、図 14 (A) は複数のサブ文書を含む XML 文書の一例を示す図、図 14 (B) は図 14 (A) に示す XML 文書の圧縮状態を示す図である。

【0142】図 14 (A) には、圧縮前つまり圧縮対象の XML 文書の一例として、同一のデータ構造をもつ複数（図中 3 つ）のサブ文書を含む、発注伝票についての XML 文書が示されている。この図 14 (A) に示す XML 文書では、開始タグ<商品>と終了タグ</商品>とで囲まれた領域が 3 つ存在し、これらの領域は全く同じデータ構造を有している。つまり、各領域においては、メーカー、製品番号および価格についてのタグと要素内容とが記述されている。ただし、これらの領域に記述された要素内容は異なっている。

【0143】図 14 (A) に示す XML 文書を、構造化文書圧縮装置 160 により圧縮する場合、サブ文書抽出部 162 において、サブ文書の抽出基準として開始タグ<商品>および終了タグ</商品>を予め設定しておくことにより、図 14 (A) に示す XML 文書から、開始タグ<商品>と終了タグ</商品>とにより囲まれた、3 つの領域がサブ文書として抽出される。

【0144】抽出された各サブ文書に対し、構造化文書圧縮装置 120 による圧縮処理を施した結果、図 14 (A) に示すサブ文書中のタグは区切りコード「,」に置き換えられる。そして、置換処理後のサブ文書とサブ文書以外の部分とをサブ文書統合部 163 により統合すると、図 14 (B) に示すような XML 圧縮文書が生成される。

【0145】一方、図 13 に示す構造化文書復元装置 260 においては、まず、例えば図 14 (B) に示すような XML 圧縮文書を復元対象として入力部 161 により取り込み、圧縮処理時と同様、サブ文書抽出部 162 により、その XML 圧縮文書から、開始タグ<商品>と終了タグ</商品>とで囲まれた領域（実質的な XML 圧縮文書）がサブ文書として抽出され、そのサブ文書（XML 圧縮文書）が構造化文書復元装置 220 に出力される。

【0146】そして、構造化文書復元装置 220 においては、入力されたサブ文書に対し、第 2 実施形態で前述した復元処理が施されて、区切りコード「,」が適切なタグに復元され、サブ文書が元の XML 文書に復元され

る。復元されたサブ文書は、サブ文書統合部 163 によりサブ文書以外の部分と統合され、統合された XML 文書が復元結果として出力部 164 から出力される。

【0147】このように、本発明の第 5 実施形態における構造化文書圧縮装置 160 によれば、一つの XML 文書中に、同一のデータ構造を有する領域（サブ文書）が複数存在する場合、XML 文書からそのサブ文書が抽出され、各サブ文書中のタグを区切りコード「,」に置き換えることにより、各サブ文書について、XML 文書の利点であるデータ構造の視認性や柔軟性／拡張性の高さを生かしたまま、XML 文書が圧縮されて XML 文書のデータ量を削減することができる。

【0148】従って、第 1 実施形態や第 2 実施形態と同様、XML 文書を格納するための記憶領域（例えば図 1、図 2、図 15～図 17 に示すハードディスク 400、410、420、440）の容量を削減することができるとともに、XML 文書データの伝送速度を高速化することができる。

【0149】また、第 5 実施形態の構造化文書復元装置 260 によれば、上述のような圧縮を施されたサブ文書を含む XML 文書が復元対象になると、その復元対象のサブ文書中で検出された区切りコード「,」を、サブ文書についてのタグリスト中のタグと対応させながら、所定のタグに置き換えるという簡易な置換処理により、復元対象の書を極めて容易に元の構造化文書に復元することができる。

【0150】〔6〕タグリストの一括管理手法の説明次に、複数種類のデータ構造（つまり複数種類のタグリスト）を一つの構造化文書処理システムで管理する場合の、本実施形態におけるタグリストの一括管理手法について、図 15～図 18 を参照しながら説明する。なお、図 15～図 17 は、それぞれ、本実施形態におけるタグリストの一括管理手法の第 1 例～第 3 例を説明するための図、図 18 本実施形態におけるタグリストの一括管理手法の第 2 例および第 3 例におけるタグリスト識別情報の付加例を示す図である。である。

【0151】図 15 に示す構造化文書処理システムでは、ハードディスク（データベース）410 において、構造化文書圧縮装置 120～160 により生成された複数（図 15 では 3 つ）の XML 圧縮文書が格納される。ここで、3 つの XML 圧縮文書には、それぞれ、識別情報（識別子）1～3 が付与されているものとする。

【0152】そして、ハードディスク 410 には、タグリスト群保持部 411 およびタグリスト管理部 412 が保持されている。タグリスト群保持部 411 は、処理対象となる XML 文書のデータ構造に対応した複数種類（図 15 では 2 種類）のタグリストを予め保持するものである。ここで、2 種類のタグリストには、それぞれ、タグリスト識別情報（タグリスト識別子）A、B が付与されているものとする。

【0153】タグリスト管理部 412 は、構造化文書圧縮装置 120～160 によって生成された XML 圧縮文書の識別情報 1～3 と、タグリスト群保持部 411 に保持されているタグリスト A、B との対応関係をテーブルによって一括管理するものである。例えば図 15 に示すタグリスト管理部 412 のテーブルによれば、XML 圧縮文書 1、2、3 とタグリスト A、A、B とがそれぞれ対応関係にある。このタグリスト管理部 412 により、ハードディスク 410 に保持されている複数の XML 圧縮文書をそれぞれ復元する際に必要なタグリストが、一括管理される。

【0154】従って、構造化文書復元装置 220、230、260 において XML 圧縮文書を復元する際には、その XML 圧縮文書の識別情報をキーにしてタグリスト管理部 412 のテーブルを検索することにより、その XML 圧縮文書の識別情報に対応した、タグリスト識別情報を得る。そして、構造化文書復元装置 210～230、260 は、そのタグリスト識別情報により特定されるタグリストを、ハードディスク 410 のタグリスト群保持部 411 から読み出し、上述したような XML 圧縮文書の復元処理に使用する。

【0155】図 16 に示す構造化文書処理システムでは、ハードディスク（データベース）420 において、構造化文書圧縮装置 110～150 により生成された複数（図 16 では 3 つ）の XML 圧縮文書 1～3 が格納されるとともに、図 15 に示したものと同様のタグリスト群保持部 411 が保持されている。

【0156】また、ハードディスク 420 には、構造化文書圧縮装置 120～160 がアクセス可能に接続されるとともに構造化文書復元装置 220、230、260 がアクセス可能に接続されており、構造化文書圧縮装置 120～160 には、タグリスト識別情報付加部 171 がそなえられるとともに、構造化文書復元装置 220、230、260 には、タグリスト識別情報取得部 172 がそなえられている。

【0157】タグリスト識別情報付加部 171 は、構造化文書圧縮装置 120～160 によって生成された XML 圧縮文書に、その XML 圧縮文書に対応するタグリストを特定するためのタグリスト識別情報を付加するものであり、タグリスト識別情報取得部 172 は、XML 圧縮文書に付加されたタグリスト識別情報を取得するものである。

【0158】従って、構造化文書圧縮装置 120～160 において XML 圧縮文書が生成されると、その XML 圧縮文書に対応するタグリスト識別情報（識別子）を、例えば図 18 に示すごとく、タグリスト識別情報付加部 171 により XML 圧縮文書の開始タグ内に属性として書き込んで付加する。なお、図 16 に示す例では、XML 圧縮文書 1～3 のそれぞれにタグリスト識別情報 A、A、B が付加されている。また、図 18 では、XML 圧

縮文書 1 または 2 における開始タグ<商品>の中に、タグリスト識別情報 A が属性「tag=「タグリスト A」」として記入された例が示されている。

【0159】一方、構造化文書復元装置 220、230、260 において XML 圧縮文書を復元する際には、その XML 圧縮文書に付加されているタグリスト識別情報を、タグリスト識別情報取得部 172 により取得する。そして、構造化文書復元装置 220、230、260 は、そのタグリスト識別情報により特定されるタグリストを、ハードディスク 420 のタグリスト群保持部 411 から読み出し、上述したような XML 圧縮文書の復元処理に使用する。

【0160】図 17 に示す構造化文書処理システムでは、管理サーバ 600 におけるハードディスク（データベース）430 に、図 15 に示したものと同様のタグリスト群保持部 411 が保持されている。また、管理サーバ 600 は、LAN 等のネットワーク 700 を介して構造化文書圧縮装置 120～160 や構造化文書復元装置 220、230、260 と通信可能に接続されるほか、これらの構造化文書圧縮装置 120～160 や構造化文書復元装置 220、230、260 は、ハードディスク（データベース）440 にアクセス可能に接続されている。このハードディスク 440 には、図 16 に示した例と同様、構造化文書圧縮装置 120～160 において生成されそれぞれタグリスト識別情報を付加された XML 圧縮文書が格納されている。

【0161】従って、構造化文書圧縮装置 120～160 において XML 圧縮文書が生成されると、図 16 に示したシステムと同様、その XML 圧縮文書に対応するタグリスト識別情報（識別子）を、例えば図 18 に示すごとく、タグリスト識別情報付加部 171 により XML 圧縮文書の開始タグ内に属性として書き込んで付加する。なお、図 17 に示す例でも、XML 圧縮文書 1～3 のそれぞれにはタグリスト識別情報 A、A、B が付加されている。

【0162】一方、構造化文書復元装置 220、230、260 において XML 圧縮文書を復元する際には、その XML 圧縮文書に付加されているタグリスト識別情報を、タグリスト識別情報取得部 172 により取得する。そして、構造化文書復元装置 220、230、260 は、そのタグリスト識別情報により特定されるタグリストを、ハードディスク 430（即ち、管理サーバ 600 上）のタグリスト群保持部 411 から、ネットワーク 700 経由で読み出し、上述したような XML 圧縮文書の復元処理に使用する。

【0163】このように、図 15～図 17 に示す構造化文書処理システムによれば、XML 圧縮文書とタグリストとの対応関係をタグリスト管理部 412 によって管理したり、XML 圧縮文書に対応するタグリストを特定するためのタグリスト識別情報を XML 圧縮文書に付加し

たりすることで、タグリスト群が一括管理され、XML 圧縮文書とタグリストとの対応関係を確実に把握することができ、XML 圧縮文書を、その XML 圧縮文書に対応したタグリストに基づいて復元することができる。

【0164】従って、XML 文書に対し圧縮・復元処理を施しながら XML 文書を取り扱うシステムにおいて、異なる種類のデータ構造（即ちタグリスト）をもつ XML 文書（XML 圧縮文書）が混在しても、混乱を招くことなく、各 XML 圧縮文書に応じたタグリストを確実に取得して復元処理を行なうことができる。

【0165】また、図 17 に示す構造化文書処理システムによれば、タグリスト群を管理サーバ 600 上で保持・管理し、ネットワーク 700 を介して処理に必要なタグリストを管理サーバ 600 から読み出すように構成することにより、タグリスト群が一括管理される。従って、複数種類のタグリストを構造化文書圧縮装置毎や構造化文書復元装置毎に管理する必要がなくなり、複数の構造化文書圧縮装置や構造化文書復元装置によって共用することができる。

【0166】なお、図 15～図 18 に示したシステムでは、タグリスト群を一括管理する場合について説明したが、タグ・属性リスト群についても上述と同様にして一括管理することができる。

【0167】〔7〕その他

なお、本発明は上述した実施形態に限定されるものではなく、本発明の趣旨を逸脱しない範囲で種々変形して実施することができる。例えば、上述した実施形態では、構造化文書が XML である場合について説明したが、本発明は、これに限定されるものではなく、タグを用いて記述される、XML と同様の構造化文書（SGML 等）に対し、上述した実施形態と同様に適用され、上述と同様の作用効果を得ることができる。

【0168】また、上述した実施形態では、区切りコードとして「」や「/」や「=」を用いた場合について説明したが、本発明は、これに限定されるものではなく、要素内容の記述に使用されることのない、他の文字あるいは記号を区切りコードとして用いてもよく、この場合も、上述した実施形態と同様の作用効果を得ることができる。

【0169】〔8〕付記

（付記 1） 同一のデータ構造を有する複数の構造化文書を圧縮する装置であって、該構造化文書におけるタグを出現順序に従って抽出してリスト化した、該複数の構造化文書について共通の一つのタグリストを取得するタグリスト取得部と、各構造化文書中のタグを所定の区切りコードに置き換えた圧縮文書を生成する構造化文書圧縮部と、該タグリスト取得部により取得された前記一つのタグリストと、該複数の構造化文書のそれぞれについて該構造化文書圧縮部により生成された複数の圧縮文書とを対応させ該複数の構造化文書の圧縮結果として出力

する出力部とをそなえたことを特徴とする、構造化文書圧縮装置。

【0170】(付記2) 該構造化文書圧縮部が、前記の各構造化文書中のタグを検出するタグ検出部と、該タグ検出部により検出された該タグを前記所定の区切りコードに置き換えて圧縮するタグ圧縮部とをそなえて構成されていることを特徴とする、付記1記載の構造化文書圧縮装置。

【0171】(付記3) 構造化文書を圧縮する装置であって、該構造化文書中のタグを検出するタグ検出部と、該タグ検出部により検出された該タグを所定の区切りコードに置き換えて圧縮するタグ圧縮部とをそなえたことを特徴とする、構造化文書圧縮装置。

【0172】(付記4) 構造化文書を圧縮する装置であって、該構造化文書から、所定の要素名をもつ開始タグと終了タグとで囲まれた領域をサブ文書として抽出するサブ文書抽出部と、該サブ文書抽出部により抽出された該サブ文書中のタグを検出するタグ検出部と、該タグ検出部により検出された該タグを所定の区切りコードに置き換えて圧縮するタグ圧縮部とをそなえたことを特徴とする、構造化文書圧縮装置。

【0173】(付記5) 該タグ検出部により検出された該タグが属性値をもつ属性付きタグであるか否かを検出する属性付きタグ検出部と、該属性付きタグ検出部により検出された該属性付きタグを前記属性値および所定の区切りコードに置き換えて圧縮する属性タグ付きタグ圧縮部とをそなえたことを特徴とする、付記3または付記4に記載の構造化文書圧縮装置。

【0174】(付記6) 所定のデータ構造を定義すべく所定の順序でタグを並べたタグリストを予め保持するタグリスト保持部と、該タグリスト保持部に保持された前記タグリストに従って、圧縮前の前記構造化文書のタグを前記所定の順序に並び替えるタグ並び替え部と、該タグリスト保持部に保持された前記タグリストに従って、該構造化文書中で省略されているタグを補完する省略タグ補完部とをそなえたことを特徴とする、付記3または付記4に記載の構造化文書圧縮装置。

【0175】(付記7) 所定のデータ構造を定義すべく所定の順序で並べたタグと属性名とをもつタグ・属性リストを予め保持するタグ・属性リスト保持部と、該タグ・属性リスト保持部に保持された前記タグ・属性リストに従って、圧縮前の前記構造化文書のタグおよび属性を前記所定の順序に並び替えるタグ・属性並び替え部と、該タグ・属性リスト保持部に保持された前記タグ・属性リストに従って、該構造化文書中で省略されているタグおよび属性を補完する省略タグ・属性補完部とをそなえたことを特徴とする、付記5記載の構造化文書圧縮装置。

【0176】(付記8) 同一のデータ構造を有する複数の構造化文書を圧縮する方法であって、該構造化文書

におけるタグを出現順序に従って抽出してリスト化した、該複数の構造化文書について共通の一つのタグリストを取得し、各構造化文書中のタグを所定の区切りコードに置き換えた圧縮文書を生成し、前記一つのタグリストと、該複数の構造化文書のそれぞれについて生成された複数の圧縮文書とを対応させ該複数の構造化文書の圧縮結果として出力することを特徴とする、構造化文書圧縮方法。

【0177】(付記9) 構造化文書を圧縮する方法であって、該構造化文書中のタグを検出し、検出された該タグを所定の区切りコードに置き換えて圧縮することを特徴とする、構造化文書圧縮方法。

【0178】(付記10) 構造化文書を圧縮する方法であって、該構造化文書から、所定の要素名をもつ開始タグと終了タグとで囲まれた領域をサブ文書として抽出し、該サブ文書中のタグを検出し、検出された該タグを所定の区切りコードに置き換えて圧縮することを特徴とする、構造化文書圧縮方法。

【0179】(付記11) 同一のデータ構造を有する複数の構造化文書を圧縮する機能をコンピュータにより実現するための構造化文書圧縮プログラムを格納したコンピュータ読取可能な記録媒体であって、該構造化文書圧縮プログラムが、該構造化文書におけるタグを出現順序に従って抽出してリスト化した、該複数の構造化文書について共通の一つのタグリストを取得するタグリスト取得部、各構造化文書中のタグを所定の区切りコードに置き換えた圧縮文書を生成する構造化文書圧縮部、および、該タグリスト取得部により取得された前記一つのタグリストと、該複数の構造化文書のそれぞれについて該構造化文書圧縮部により生成された複数の圧縮文書とを対応させ該複数の構造化文書の圧縮結果として出力する出力部として、該コンピュータに機能させることを特徴とする、構造化文書圧縮プログラムを格納したコンピュータ読取可能な記録媒体。

【0180】(付記12) 構造化文書を圧縮する機能をコンピュータにより実現するための構造化文書圧縮プログラムを格納したコンピュータ読取可能な記録媒体であって、該構造化文書圧縮プログラムが、該構造化文書中のタグを検出するタグ検出部、および、該タグ検出部により検出された該タグを所定の区切りコードに置き換えて圧縮するタグ圧縮部として、該コンピュータを機能させることを特徴とする、構造化文書圧縮プログラムを格納したコンピュータ読取可能な記録媒体。

【0181】(付記13) 構造化文書を圧縮する機能をコンピュータにより実現するための構造化文書圧縮プログラムを格納したコンピュータ読取可能な記録媒体であって、該構造化文書圧縮プログラムが、該構造化文書から、所定の要素名をもつ開始タグと終了タグとで囲まれた領域をサブ文書として抽出するサブ文書抽出部、該サブ文書抽出部により抽出された該サブ文書中のタグを

検出するタグ検出部、および、該タグ検出部により検出された該タグを所定の区切りコードに置き換えて圧縮するタグ圧縮部として、該コンピュータを機能させることを特徴とする、構造化文書圧縮プログラムを格納したコンピュータ読取可能な記録媒体。

【0182】（付記14） 同一のデータ構造を有する複数の構造化文書中のタグを所定の区切りコードに置き換えることにより生成された複数の圧縮文書を、該複数の構造化文書におけるタグを出現順序に従ってリスト化したタグリストに基づいて復元する装置であって、該タグリストに対応するデータ構造をメモリ上に複製データ構造として展開・複製する複製部と、該複製データ構造におけるタグの位置と各圧縮文書中の前記所定の区切りコードの位置とを対応させながら、各圧縮文書中の要素内容を該メモリ上における該複製データ構造の所定領域に書き出す書出部とをそなえたことを特徴とする、構造化文書復元装置。

【0183】（付記15） 構造化文書中のタグを所定の区切りコードに置き換えることにより生成された圧縮文書を復元する装置であって、該構造化文書におけるタグを出現順序に従ってリスト化したタグリストを予め保持するタグリスト保持部と、該圧縮文書中の前記所定の区切りコードを検出する区切りコード検出部と、該タグリストにおけるタグの位置と該区切りコード検出部により検出された前記所定の区切りコードの位置とを対応させながら、該区切りコード検出部により検出された前記所定の区切りコードを、該タグリストにおける対応するタグに置き換えて復元するタグ復元部とをそなえたことを特徴とする、構造化文書復元装置。

【0184】（付記16） 構造化文書において所定の要素名をもつ開始タグと終了タグとで囲まれた領域であるサブ文書中のタグを所定の区切りコードに置き換えることにより生成された圧縮文書を復元する装置であって、該サブ文書におけるタグを出現順序に従ってリスト化したタグリストを予め保持するタグリスト保持部と、該圧縮文書から該サブ文書を抽出するサブ文書抽出部と、該サブ文書抽出部により抽出された該サブ文書中の前記所定の区切りコードを検出する区切りコード検出部と、該タグリストにおけるタグの位置と該区切りコード検出部により検出された前記所定の区切りコードの位置とを対応させながら、該区切りコード検出部により検出された前記所定の区切りコードを、該タグリストにおける対応するタグに置き換えて復元するタグ復元部とをそなえたことを特徴とする、構造化文書復元装置。

【0185】（付記17） 該圧縮文書中において、属性付きタグ内の属性が属性値および所定の区切りコードに置き換えられて圧縮されている場合、該圧縮文書における属性名を出現順序に従ってリスト化した属性リストを予め保持する属性リスト保持部と、該タグ復元部で復元対象となったタグが属性付きタグに復元されるべきも

のであるか否かを検出する属性付きタグ検出部と、該属性付きタグについての属性値と該属性リストにおける属性名とを対応させて、該属性付きタグ検出部により検出された該属性付きタグ内の該属性を復元する属性付きタグ復元部とをそなえたことを特徴とする、付記15または付記16に記載の構造化文書復元装置。

【0186】（付記18） 同一のデータ構造を有する複数の構造化文書中のタグを所定の区切りコードに置き換えることにより生成された複数の圧縮文書を、該複数の構造化文書におけるタグを出現順序に従ってリスト化したタグリストに基づいて復元する方法であって、該タグリストに対応するデータ構造をメモリ上に複製データ構造として展開・複製し、該複製データ構造におけるタグの位置と各圧縮文書中の前記所定の区切りコードの位置とを対応させながら、各圧縮文書中の要素内容を該メモリ上における該複製データ構造の所定領域に書き出すことを特徴とする、構造化文書復元方法。

【0187】（付記19） 構造化文書中のタグを所定の区切りコードに置き換えることにより生成された圧縮文書を復元する方法であって、該構造化文書におけるタグを出現順序に従ってリスト化したタグリストを予め保持し、該圧縮文書中の前記所定の区切りコードを検出し、検出された前記所定の区切りコードの位置と該タグリストにおけるタグの位置とを対応させながら、検出された前記所定の区切りコードを該タグリストにおける対応するタグに置き換えて復元することを特徴とする、構造化文書復元方法。

【0188】（付記20） 構造化文書において所定の要素名をもつ開始タグと終了タグとで囲まれた領域であるサブ文書中のタグを所定の区切りコードに置き換えることにより生成された圧縮文書を復元する方法であって、該サブ文書におけるタグを出現順序に従ってリスト化したタグリストを予め保持し、該圧縮文書から該サブ文書を抽出し、抽出された該サブ文書中の前記所定の区切りコードを検出し、検出された前記所定の区切りコードの位置と該タグリストにおけるタグの位置とを対応させながら、検出された前記所定の区切りコードを該タグリストにおける対応するタグに置き換えて復元することを特徴とする、構造化文書復元方法。

【0189】（付記21） 同一のデータ構造を有する複数の構造化文書中のタグを所定の区切りコードに置き換えることにより生成された複数の圧縮文書を、該複数の構造化文書におけるタグを出現順序に従ってリスト化したタグリストに基づいて復元する機能をコンピュータにより実現するための構造化文書復元プログラムを格納したコンピュータ読取可能な記録媒体であって、該構造化文書復元プログラムが、該タグリストに対応するデータ構造をメモリ上に複製データ構造として展開・複製する複製部、および、該複製データ構造におけるタグの位置と各圧縮文書中の前記所定の区切りコードの位置とを

対応させながら、各圧縮文書中の要素内容を該メモリ上における該複製データ構造の所定領域に書き出す書出部として、該コンピュータを機能させることを特徴とする、構造化文書復元プログラムを格納したコンピュータ読取可能な記録媒体。

【0190】(付記22) 構造化文書中のタグを所定の区切りコードに置き換えることにより生成された圧縮文書を復元する機能をコンピュータにより実現するための構造化文書復元プログラムを格納したコンピュータ読取可能な記録媒体であって、該構造化文書復元プログラムが、該圧縮文書中の前記所定の区切りコードを検出する区切りコード検出部、および、該構造化文書におけるタグを出現順序に従ってリスト化したタグリストにおけるタグの位置と、該区切りコード検出部により検出された前記所定の区切りコードの位置とを対応させながら、該区切りコード検出部により検出された前記所定の区切りコードを、該タグリストにおける対応するタグに置き換えて復元するタグ復元部として、該コンピュータを機能させることを特徴とする、構造化文書復元プログラムを格納したコンピュータ読取可能な記録媒体。

【0191】(付記23) 構造化文書において所定の要素名をもつ開始タグと終了タグとで囲まれた領域であるサブ文書中のタグを所定の区切りコードに置き換えることにより生成された圧縮文書を復元する機能をコンピュータにより実現するための構造化文書復元プログラムを格納したコンピュータ読取可能な記録媒体であって、該構造化文書復元プログラムが、該圧縮文書から該サブ文書を抽出するサブ文書抽出部、該サブ文書抽出部により抽出された該サブ文書中の前記所定の区切りコードを検出する区切りコード検出部、および、該サブ文書におけるタグを出現順序に従ってリスト化したタグリストにおけるタグの位置と、該区切りコード検出部により検出された前記所定の区切りコードの位置とを対応させながら、該区切りコード検出部により検出された前記所定の区切りコードを、該タグリストにおける対応するタグに置き換えて復元するタグ復元部として、該コンピュータを機能させることを特徴とする、構造化文書復元プログラムを格納したコンピュータ読取可能な記録媒体。

【0192】(付記24) 同一のデータ構造を有する複数の構造化文書に対する処理を行なうべく、該複数の構造化文書を圧縮する構造化文書圧縮装置と、該構造化文書圧縮装置による圧縮データを該複数の構造化文書に復元する構造化文書復元装置とを含んで構成される構造化文書処理システムにおいて、該構造化文書圧縮装置が、該構造化文書におけるタグを出現順序に従って抽出してリスト化した、該複数の構造化文書について共通の一つのタグリストを取得するタグリスト取得部と、各構造化文書中のタグを所定の区切りコードに置き換えた圧縮文書を生成する構造化文書圧縮部と、該タグリスト取得部により取得された前記一つのタグリストと、該複数の

の構造化文書のそれぞれについて該構造化文書圧縮部により生成された複数の圧縮文書とを対応させ該複数の構造化文書の圧縮結果として出力する出力部とをそなえて構成されるとともに、該構造化文書復元装置が、該複数の圧縮文書の復元結果を格納するメモリと、該タグリストに対応するデータ構造を該メモリ上に複製データ構造として展開・複製する複製部と、該複製データ構造におけるタグの位置と各圧縮文書中の前記所定の区切りコードの位置とを対応させながら、各圧縮文書中の要素内容を該メモリ上における該複製データ構造の所定領域に書き出す書出部とをそなえて構成されたことを特徴とする、構造化文書処理システム。

【0193】(付記25) 構造化文書に対する処理を行なうべく、該構造化文書を圧縮する構造化文書圧縮装置と、該構造化文書圧縮装置による圧縮データを該構造化文書に復元する構造化文書復元装置とを含んで構成される構造化文書処理システムにおいて、該構造化文書圧縮装置が、該構造化文書中のタグを検出するタグ検出部と、該タグ検出部により検出された該タグを所定の区切りコードに置き換えて圧縮するタグ圧縮部とをそなえて構成されるとともに、該構造化文書復元装置が、該構造化文書におけるタグを出現順序に従ってリスト化したタグリストを予め保持するタグリスト保持部と、該圧縮文書中の前記所定の区切りコードを検出する区切りコード検出部と、該タグリストにおけるタグの位置と該区切りコード検出部により検出された前記所定の区切りコードの位置とを対応させながら、該区切りコード検出部により検出された前記所定の区切りコードを、該タグリストにおける対応するタグに置き換えて復元するタグ復元部とをそなえて構成されたことを特徴とする、構造化文書処理システム。

【0194】(付記26) 構造化文書に対する処理を行なうべく、該構造化文書を圧縮する構造化文書圧縮装置と、該構造化文書圧縮装置による圧縮データを該構造化文書に復元する構造化文書復元装置とを含んで構成される構造化文書処理システムにおいて、該構造化文書圧縮装置が、該構造化文書から、所定の要素名をもつ開始タグと終了タグとで囲まれた領域をサブ文書として抽出するサブ文書抽出部と、該サブ文書抽出部により抽出された該サブ文書中のタグを検出するタグ検出部と、該タグ検出部により検出された該タグを所定の区切りコードに置き換えて圧縮するタグ圧縮部とをそなえて構成されるとともに、構造化文書復元装置が、該サブ文書におけるタグを出現順序に従ってリスト化したタグリストを予め保持するタグリスト保持部と、該圧縮文書から該サブ文書を抽出するサブ文書抽出部と、該サブ文書抽出部により抽出された該サブ文書中の前記所定の区切りコードを検出する区切りコード検出部と、該タグリストにおけるタグの位置と該区切りコード検出部により検出された前記所定の区切りコードの位置とを対応させながら、該

区切りコード検出部により検出された前記所定の区切りコードを、該タグリストにおける対応するタグに置き換えて復元するタグ復元部とをそなえて構成されたことを特徴とする、構造化文書処理システム。

【0195】（付記27） 処理対象となりうる構造化文書のデータ構造に対応した複数のタグリストを予め保持するタグリスト群保持部と、該構造化文書圧縮装置によって生成された該圧縮文書と、該タグリスト群保持部に保持されている該タグリストとの対応関係を管理するタグリスト管理部とをそなえたことを特徴とする、付記25または付記26に記載の構造化文書処理システム。

【0196】（付記28） 処理対象となりうる構造化文書のデータ構造に対応した複数のタグリストを予め保持するタグリスト群保持部と、該構造化文書圧縮装置によって生成された該圧縮文書に、該圧縮文書に対応するタグリストを特定するためのタグリスト識別情報を付加するタグリスト識別情報付加部と、該圧縮文書に付加された前記タグリスト識別情報を取得するタグリスト識別情報取得部とをそなえ、該構造化文書復元装置が、該タグリスト識別情報取得部によって取得された前記タグリスト識別情報に対応する該タグリストを用いて、該圧縮文書を復元することを特徴とする、付記25または付記26に記載の構造化文書処理システム。

【0197】（付記29） 該タグリスト群保持部が管理サーバ上に配置され、処理に必要なタグリストが、ネットワークを介して該管理サーバ上の該タグリスト群保持部から読み出されることを特徴とする、付記27または付記28に記載の構造化文書処理システム。

【0198】

【発明の効果】以上詳述したように、本発明の構造化文書圧縮装置（請求項1、2）および構造化文書復元装置（請求項3、4）並びに構造化文書処理システム（請求項5）によれば、以下のような効果ないし利点を得ることができる。

（1）本発明により生成される圧縮文書では、タグが所定の区切りコードに置換されているだけで、データ内容（要素内容）はそのまま記述されているので、構造化文書の利点であるデータ構造の視認性や柔軟性／拡張性を生かしたまま、構造化文書を圧縮して構造化文書のデータ量を削減することができる。従って、構造化文書を格納するための記憶領域の容量を削減できるとともに構造化文書データの伝送速度を高速化することができる（請求項1、5）。

【0199】（2）複数の構造化文書の圧縮結果は、データ構造（一つのタグリスト）とデータ内容（複数の圧縮文書）とに分離されているので、一つのタグリストに対する解析処理を一度だけ行なって、複数の圧縮文書に共通のデータ構造を取得してしまえば、後は、取得されたデータ構造を複製して流用することにより、圧縮文書毎に一々タグ解析を行なう必要がなくなる。従って、同

一のデータ構造を有する多数の構造化文書を取り扱う際に、無駄なタグ解析を行なう必要が一切なくなり、タグ解析の負荷が大幅に低減され、構造化文書をメモリに展開する際の処理速度を飛躍的に高速化することができる（請求項1、3、5）。

【0200】（3）構造化文書中で検出されたタグを所定の区切りコードに置換するという極めて単純な置換処理によって、構造化文書の利点であるデータ構造の視認性や柔軟性／拡張性の高さを生かしたまま、構造化文書を圧縮して構造化文書のデータ量を削減することができる。従って、構造化文書を格納するための記憶領域の容量を削減できるとともに構造化文書データの伝送速度を高速化することができる（請求項2）。このような圧縮を行なった場合、圧縮文書中で検出された所定の区切りコードを、その圧縮文書についてのタグリスト中のタグと対応させながら、所定のタグに置き換えるという簡易な置換処理によって、圧縮文書を極めて容易に元の構造化文書に復元することができる（請求項4）。

【0201】（4）一つの構造化文書中に、同一のデータ構造を有する領域（サブ文書）が複数存在する場合、構造化文書から、そのサブ文書が、所定の要素名をもつ開始タグと終了タグとで囲まれた領域として抽出され、各サブ文書中のタグを所定の区切りコードに置き換えることにより、各サブ文書について、構造化文書の利点であるデータ構造の視認性や柔軟性／拡張性の高さを生かしたまま、構造化文書を圧縮して構造化文書のデータ量を削減することができる。従って、構造化文書を格納するための記憶領域の容量を削減できるとともに構造化文書データの伝送速度を高速化することができる。このような圧縮を行なった場合、復元対象におけるサブ文書中で検出された所定の区切りコードを、サブ文書についてのタグリスト中のタグと対応させながら、所定のタグに置き換えるという簡易な置換処理によって、復元対象の文書を極めて容易に元の構造化文書に復元することができる。

【0202】（5）タグが属性値をもつ属性付きタグである場合には、その属性付きタグを属性値および所定の区切りコードに置換えて圧縮する。これにより、圧縮文書において属性値がそのまま記述されるので、属性値の視認性を保ちながら構造化文書の圧縮を行なうことができる。このような圧縮を施された属性付きタグが復元対象になると、その属性付きタグについての属性値とその圧縮文書についての属性リスト中の属性名とを対応させることにより、属性付きタグを極めて容易に復元することができる。

【0203】（6）所定のデータ構造を定義する、タグリストまたはタグ・属性リストに従って、圧縮前の前記構造化文書のタグまたは属性を所定の順序に並び替えるとともに、構造化文書中で省略されているタグまたは属

性を補完することにより、タグまたは属性の記述順序の逆転や、タグまたは属性の記述の欠落といった不備をもつ構造化文書は、所定のデータ構造を有するように正規化される。従って、同一のデータ構造を有する多数の構造化文書を圧縮処理対象とする場合、上述のような不備をもつ構造化文書が含まれていても、圧縮処理前に、圧縮処理対象の全ての構造化文書が、タグリストまたはタグ・属性リストで定義された所定のデータ構造を有するように正規化される。これにより、多数の構造化文書（圧縮文書）を、一つのタグリストまたはタグ・属性リストによって確実に管理し取り扱うことが可能になる。

【0204】（7）圧縮文書とタグリストとの対応関係をタグリスト管理部によって管理したり、圧縮文書に対応するタグリストを特定するためのタグリスト識別情報を圧縮文書に付加したりすることで、タグリスト群が一括され、圧縮文書とタグリストとの対応関係を確実に把握でき、圧縮文書を、その圧縮文書に対応したタグリストに基づいて復元することができる。従って、構造化文書に対し圧縮・復元処理を施しながら構造化文書を取り扱うシステムにおいて、異なる種類のデータ構造（即ちタグリスト）をもつ構造化文書（圧縮文書）が混在しても、混乱を招くことなく、各圧縮文書に応じたタグリストを確実に取得して復元処理を行なうことができる。

【0205】（8）タグリスト群を管理サーバ上で保持・管理し、ネットワークを介して処理に必要なタグリストを管理サーバから読み出すように構成することにより、タグリスト群が一括管理される。従って、複数種類のタグリストを圧縮装置毎や復元装置毎に管理する必要がなくなり、複数の圧縮装置や復元装置によって共用することができる。

【図面の簡単な説明】

【図1】本発明の第1実施形態における構造化文書圧縮装置の構成を示すブロック図である。

【図2】本発明の第1実施形態における構造化文書復元装置（メモリ展開部）の構成を示すブロック図である。

【図3】（A）～（C）はいずれも第1実施形態におけるデータ例を示すもので、（A）はXML文書の一例を示す図、（B）は（A）に示すXML文書から得られたタグリストを示す図、（C）は（A）に示すXML文書の圧縮状態を示す図である。

【図4】本発明の第2実施形態における構造化文書圧縮装置の構成を示すブロック図である。

【図5】本発明の第2実施形態における構造化文書復元装置の構成を示すブロック図である。

【図6】（A）～（D）はいずれも第2実施形態におけるデータ例を示すもので、（A）はXML文書の一例を示す図、（B）は（A）に示すXML文書に対応するタグリストを示す図、（C）は（A）に示すXML文書の圧縮状態の一例を示す図、（D）は（A）に示すXML文書の圧縮状態の他例を示す図である。

【図7】本発明の第3実施形態における構造化文書圧縮装置の構成を示すブロック図である。

【図8】本発明の第3実施形態における構造化文書復元装置の構成を示すブロック図である。

【図9】（A）～（D）はいずれも第3実施形態におけるデータ例を示すもので、（A）はXML文書の一例を示す図、（B）は（A）に示すXML文書に対応するタグ・属性リストを示す図、（C）は（A）に示すXML文書の圧縮状態の一例を示す図、（D）は（A）に示すXML文書の圧縮状態の他例を示す図である。

【図10】本発明の第4実施形態における構造化文書圧縮装置の要部構成を示すブロック図である。

【図11】（A）～（C）はいずれも第4実施形態におけるデータ例を示すもので、（A）はタグリストの一例を示す図、（B）はタグの記述に不備のあるXML文書の一例を示す図、（C）は（B）に示すXML文書を（A）に示すタグリストに従って正規化した結果を示す図である。

【図12】本発明の第4実施形態における構造化文書圧縮装置の変形例の要部構成を示すブロック図である。

【図13】本発明の第5実施形態における構造化文書圧縮装置および構造化文書復元装置の構成を示すブロック図である。

【図14】（A）および（B）はいずれも第5実施形態におけるデータ例を示すもので、（A）は複数のサブ文書を含むXML文書の一例を示す図、（B）は（A）に示すXML文書の圧縮状態を示す図である。

【図15】本実施形態におけるタグリストの一括管理手法の第1例を説明するための図である。

【図16】本実施形態におけるタグリストの一括管理手法の第2例を説明するための図である。

【図17】本実施形態におけるタグリストの一括管理手法の第3例を説明するための図である。

【図18】本実施形態におけるタグリストの一括管理手法の第2例および第3例におけるタグリスト識別情報の付加例を示す図である。

【図19】一般的なユニバーサルデータ圧縮について説明するための図である。

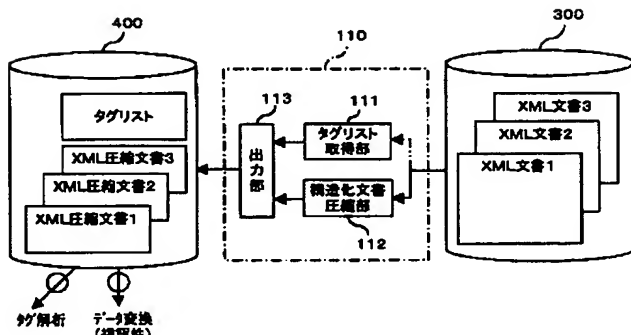
【図20】XML文書を取り扱う一般的なシステムの構成例を示すブロック図である。

【符号の説明】

- 110 構造化文書圧縮装置
- 111 タグリスト取得部
- 112 構造化文書圧縮部
- 113 出力部
- 120 構造化文書圧縮装置
- 121 入力部
- 122 タグ検出部
- 123 タグ圧縮部
- 124 出力部

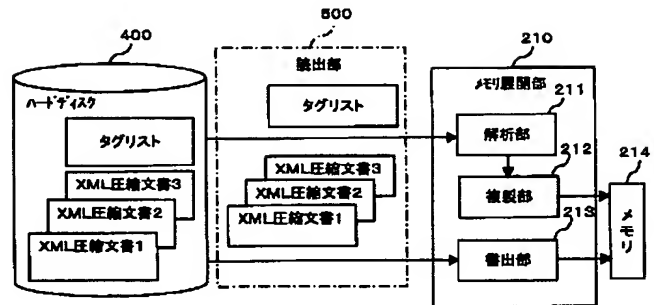
- 130 構造化文書圧縮装置
- 131 属性付きタグ検出部
- 132 属性付きタグ圧縮部
- 140 構造化文書圧縮装置
- 141 入力部
- 142 タグリスト保持部
- 143 タグ並び替え部
- 144 省略タグ補完部
- 150 構造化文書圧縮装置
- 151 入力部
- 152 タグ・属性リスト保持部
- 153 タグ・属性並び替え部
- 154 省略タグ・属性補完部
- 160 構造化文書圧縮装置
- 161 入力部
- 162 サブ文書抽出部
- 163 サブ文書統合部
- 164 出力部
- 171 タグリスト識別情報付加部
- 172 タグリスト識別情報取得部
- 210 メモリ展開部 (構造化文書復元装置)
- 211 解析部

【図 1】



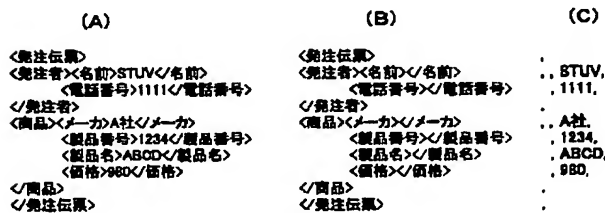
- 212 複製部
- 213 書出部
- 214 メモリ
- 220 構造化文書復元装置
- 221 入力部
- 222 タグリスト保持部
- 223 区切りコード検出部
- 224 タグ復元部
- 225 出力部
- 10 230 構造化文書復元装置
- 231 属性リスト保持部
- 232 属性付きタグ検出部
- 233 属性付きタグ復元部
- 260 構造化文書復元装置
- 300, 400, 410, 420, 430, 440 ハードディスク (データベース)
- 411 タグリスト群保持部
- 412 タグリスト管理部
- 500 読出部
- 20 600 管理サーバ
- 700 ネットワーク

【図 2】



【図 5】

【図 3】

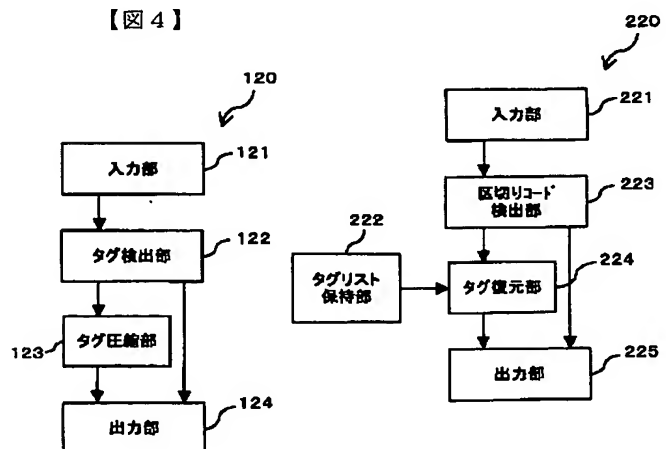


【図 18】

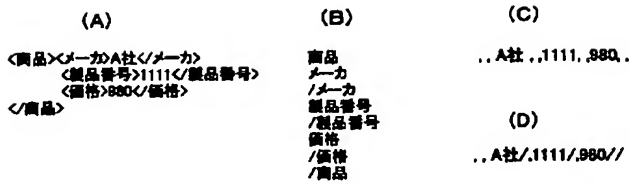
<商品 tag="221" A">
A社, 111A, 990,
</商品>

識別子

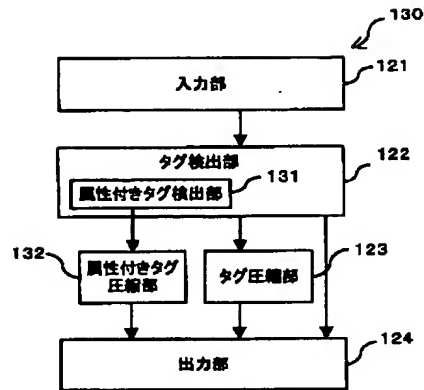
【図 4】



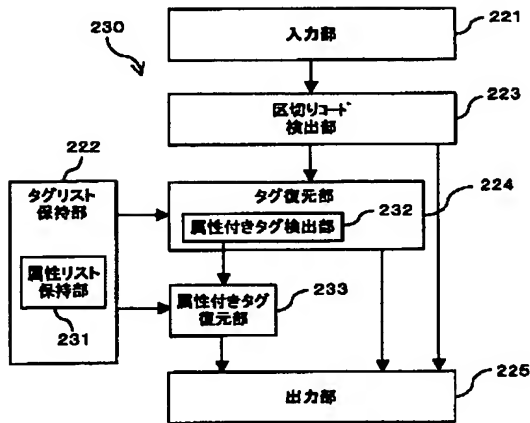
【図 6】



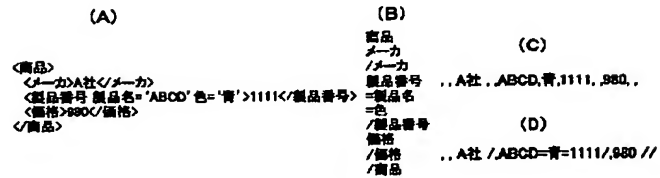
【図 7】



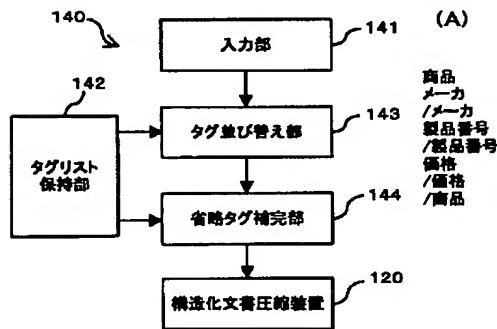
【図 8】



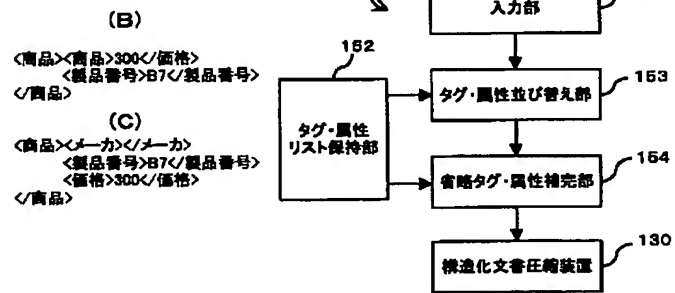
【図 9】



【図 10】

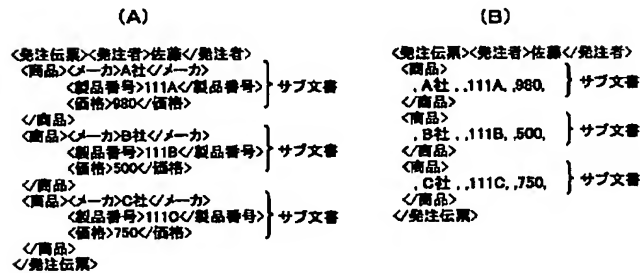


【図 11】

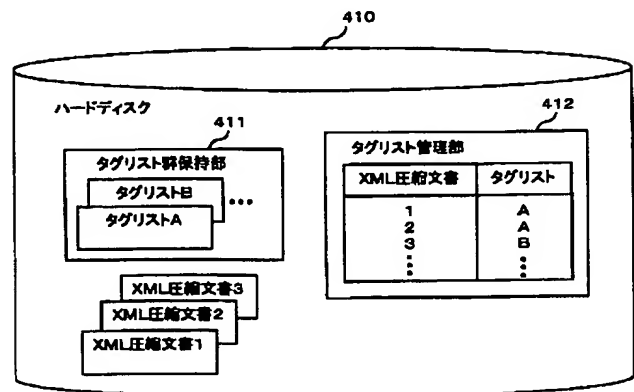


【図 12】

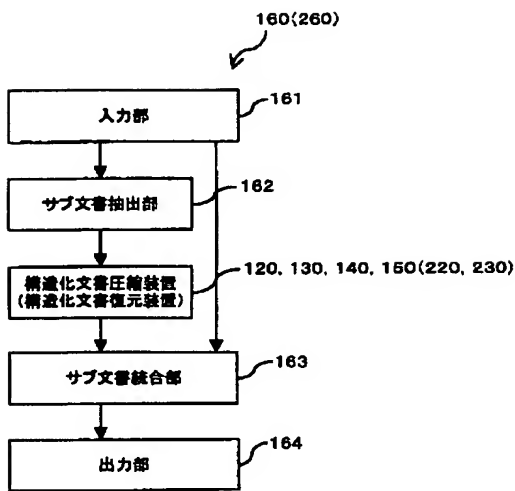
【図 14】



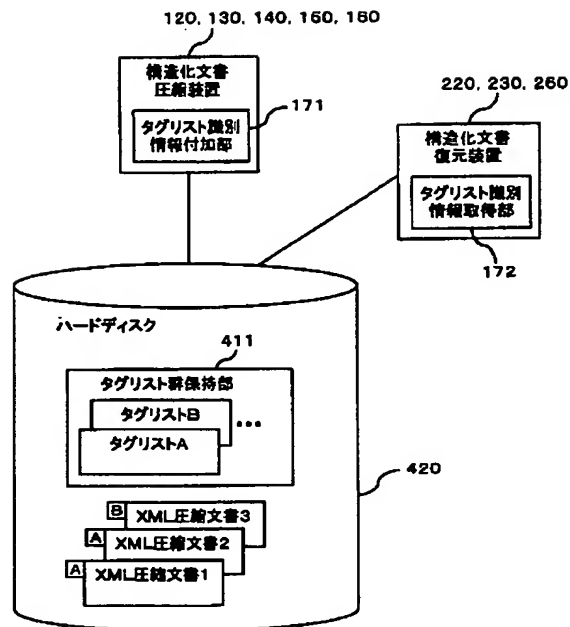
【図 15】



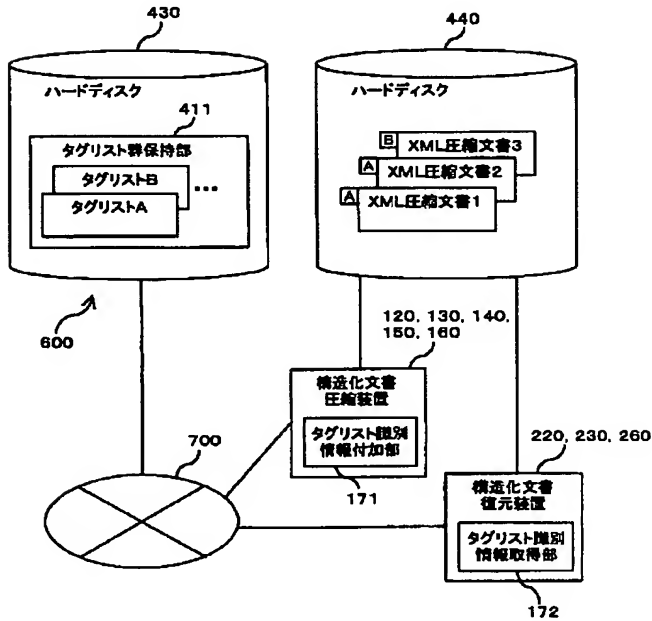
【図13】



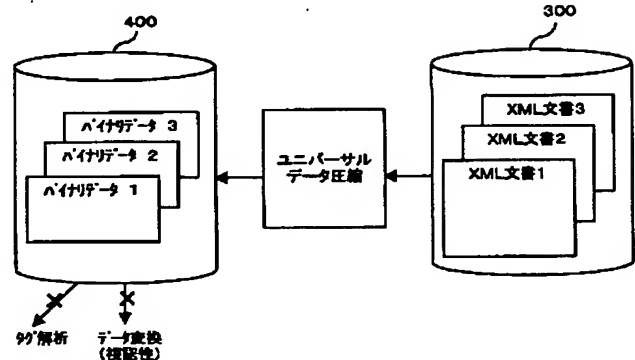
【図16】



【図17】



【図19】



【図20】

